Федеральный исследовательский центр «Информатика и управление» РАН
Московский государственный университет имени М.В. Ломоносова
Национальный исследовательский ядерный университет «МИФИ»
Московская секция ACM SIGMOD
Российский фонд фундаментальных исследований

АНАЛИТИКА И УПРАВЛЕНИЕ ДАННЫМИ В ОБЛАСТЯХ С ИНТЕНСИВНЫМ ИСПОЛЬЗОВАНИЕМ ДАННЫХ

Сборник научных трудов XIX Международной конференции
DAMDID / RCDL'2017

10-13 октября 2017 г.
г. Москва, МГУ, Россия

Под редакцией Л.А. Калиниченко, Я. Манолопулос, Н.А. Скворцова, В.А. Сухомлина

Federal Research Center "Computer Science and Control" of RAS
Lomonosov Moscow State University
Institute for Nuclear Power Engineering MEPhI
Moscow ACM SIGMOD Chapter
Russian Foundation for Basic Research

DATA ANALYTICS AND MANAGEMENT IN DATA INTENSIVE DOMAINS

Collection of Scientific Papers of the XIX International Conference
DAMDID / RCDL'2017

October 10-13, 2017
Moscow, MSU, Russia

Edited by L.A. Kalinichenko, Y. Manolopoulos, N.A. Skvortsov, V.A. Sukhomlin

Москва — 2017

УДК [002:004.9] (063)
ББК [73+32.973.233]я431
А 64

Издание осуществлено при финансовой поддержке Российского фонда фундаментальных исследований
(Проект № 17–07–20546 г)

А 64 Аналитика и управление данными в областях с интенсивным использованием данных [Текст]: сборник научных трудов XIX Международной конференции DAMDID / RCDL'2017 (10–13 октября 2017 года, г. Москва, МГУ, Россия) / Под ред. Л. А. Калиниченко, Я. Манолопулос, Н. А. Скворцова, В. А. Сухомлина. — Москва: ФИЦ ИУ РАН, 2017. — 498 с.

ISBN 978–5–519–60516–8

Конференция «Аналитика и управление данными в областях с интенсивным использованием данных» («Data Analytics and Management in Data Intensive Domains») представляет собой мультидисциплинарный форум исследователей и практиков из разнообразных областей деятельности людей, содействующий сотрудничеству и обмену идеями в сфере анализа и управления данными в областях исследований, движимых интенсивным использованием данных (ОИИД). Подходы к анализу данных и управлению данными, развиваемые в конкретных ОИИД X-информатики (таких как X=астро, био, гео, нейро, мед, физика, химия, и пр.), социальных наук, а также информатики, различных ОИИД промышленности, новых технологий, финансов и бизнеса составляют предметную область конференции. Конференция DAMDID была образована в 2015 г. в результате трансформации конференции RCDL («Электронные библиотеки: перспективные методы и технологии, электронные коллекции») с сохранением преемственности по отношению к RCDL после многих лет ее успешного функционирования.

ББК [73+32.973.233]я431

The publication was made with the financial support of the Russian Foundation for Basic Research
(Project № 17–07–20546 g)

Data Analytics and Management in Data Intensive Domains: Collection of Scientific Papers of the XIX International Conference DAMDID / RCDL'2017 (October 10–13, 2017, Moscow, Russia). Edited by L.A. Kalinichenko, Y. Manolopoulos, N.A. Skvortsov, V.A. Sukhomlin. — Moscow: FRC CSC RAS, 2017. — 498 p.

ISBN 978–5–519–60516–8

The «Data Analytics and Management in Data Intensive Domains» conference (DAMDID) traditionally is planned as a multidisciplinary forum of researchers and practitioners from various domains of science and research, promoting cooperation and exchange of ideas in the area of data analysis and management in domains driven by data intensive research. Approaches to data analysis and management being developed in specific data intensive domains (DID) of X-informatics (such as X =astro, bio, chemo, geo, med, neuro, physics, etc.), social sciences, as well as in various branches of informatics, industry, new technologies, finance and business constitute the universe of the conference discourse. DAMDID conference was formed in 2015 as a result of transformation of the RCDL conference («Digital libraries: advanced methods and technologies, digital collections», http://rcdl.ru) so that the continuity with RCDL has been preserved after many years of its successful work.

ISBN 978-5-519-60516-8

XIX Международная конференция DAMDID / RCDL'2017

АНАЛИТИКА И УПРАВЛЕНИЕ ДАННЫМИ В ОБЛАСТЯХ С ИНТЕНСИВНЫМ ИСПОЛЬЗОВАНИЕМ ДАННЫХ

10—13 октября 2017 года, г. Москва
damdid2017.frccsc.ru

КООРДИНАЦИОННЫЙ КОМИТЕТ

Сопредседатели

Колчанов Николай Александрович,
академик РАН (ИЦиГ СО РАН, Новосибирск)

Соколов Игорь Анатольевич,
академик РАН (ФИЦ ИУ РАН, Москва)

Заместитель председателя

Калиниченко Леонид Андреевич
(ФИЦ ИУ РАН, Москва)

Члены координационного комитета

Авраменко Аркадий Ефимович
(Радиоастрономическая обсерватория, Пущино)

Браславский Павел Исаакович
(Уральский федеральный университет, Екатеринбург)

Бунаков Василий Эрикович
(STFC, Харвелл, Оксфордшир, Великобритания)

Вольфенгаген Вячеслав Эрнстович
(МИФИ, Москва)

Воронцов Константин Вячеславович
(МГУ, Москва)

Елизаров Александр Михайлович
(Казанский федеральный университет, Казань)

Захаров Виктор Николаевич
(ФИЦ ИУ РАН, Москва)

Климентов Алексей Анатольевич
(Brookhaven National Laboratory, США)

Когаловский Михаил Рувимович
(ИПР РАН, Москва)

Кореньков Владимир Васильевич
(ОИЯИ, Дубна)

Кузнецов Сергей Дмитриевич
(ИСП РАН, Москва)

Кузьминский Михаил Борисович
(ИОХ РАН, Москва)

Литвин Владимир Андреевич
(Evogh Inc, CalTech, США)

Майсурадзе Арчил Ивериевич
(МГУ, Москва)

Малков Олег Юрьевич
(ИНАСАН, Москва)

Марчук Александр Гурьевич
(ИСИ СО РАН, Новосибирск)

Некрестьянов Игорь Сергеевич
(Verizon, США)

Новиков Борис Асенович
(СПбГУ, Санкт Петербург)

Подколодный Николай Леонтьевич
(ИЦиГ СО РАН, Новосибирск)

Позаненко Алексей Степанович
(ИКИ РАН, Москва)

Серебряков Владимир Алексеевич
(ВЦ РАН, Москва)

Сметанин Юрий Геннадиевич
(РФФИ, Москва)

Смирнов Владимир Николаевич
(ЯрГУ, Ярославль)

Ступников Сергей Александрович
(ФИЦ ИУ РАН, Москва)

Фазлиев Александр Зарипович
(ИОА СО РАН, Томск)

Генеральный председатель конференции DAMDID/RCDL'2017

Соколов Игорь Анатольевич,
академик РАН (ФИЦ ИУ РАН, Москва)

ОРГАНИЗАЦИОННЫЙ КОМИТЕТ

Сопредседатели

Владимир Сухомлин
(ВМК МГУ, Москва)

Виктор Захаров
(ФИЦ ИУ РАН, Москва)

Александр Невзоров
(НИЯУ МИФИ, Москва)

Члены оргкомитета

Дмитрий Брюхов
(ФИЦ ИУ РАН), поддержка веб-сайта

Николай Скворцов
(ФИЦ ИУ РАН), поддержка сайта СМТ

Дмитрий Намиот *(ВМК МГУ),*

Елена Зубарева (*ВМК МГУ, ЕГУ им. И.А. Бунина),*

Николай Скворцов (*ФИЦ ИУ РАН),*

Александр Елизаров *(КФУ), взаимодействие с членами ПК и авторами в процессе сбора, рецензирования, отбора и редактирования работ для издания локального тома трудов*

Елена Зубарева
(ВМК МГУ, ЕГУ им. И.А. Бунина), техническое редактирование и типографское издание материалов конференции (программа и локальный сборник статей)

Евгений Морковин
(ВМК МГУ), привлечение участников, реклама конференции

Дмитрий Ковалев
(ФИЦ ИУ РАН), контакты, переписка

Евгения Дударева
(ФИЦ ИУ РАН), юридические и финансовые вопросы

Ирина Карзалова
(ФИЦ ИУ РАН), финансовые вопросы

Юлия Трусова
(ФИЦ ИУ РАН), визовая поддержка

Елена Зубарева
(ВМК МГУ, ЕГУ им. И.А. Бунина), техническое редактирование и типографское издание материалов конференции (программа и локальный сборник статей), председатель локального обустройства участников конференции, включая регистрацию, участников, организацию обедов и кофейных перерывов, организацию welcome party и conference dinner, размещение участников для проживания в ГЗ МГУ, организация экскурсий по МГУ и в Москве

Алексей Лифарь
(НИЯУ МИФИ), финансовая поддержка приглашенных зарубежных докладчиков, организация концерта хора МИФИ для участников конференции, организация встреч в аэропортах, проводов, транспорта для участников конференции, организация синхронного перевода

Владимир Сухомлин, Евгений Морковин
(ВМК МГУ), организация сбора регистрационных взносов

Евгений Ильюшин, Дмитрий Гурьев, Владимир Романов
(ВМК МГУ), ответственные за техническую поддержку аппаратуры (персональных компьютеров и проекторов в аудиториях), Wi-Fi в аудиториях и жилых помещениях МГУ, аппаратура для синхронного перевода

ПРОГРАММНЫЙ КОМИТЕТ

Сопредседатели

Леонид Андреевич Калиниченко
(ФИЦ ИУ РАН, Москва)

Яннис Манолопулос
(Университет им. Аристотеля в Салониках, Греция)

Дмитрий Евгеньевич Намиот
(ВМК МГУ, Москва)

Члены программного комитета

Карл Аберер
(EPFL, Лозанна, Швейцария)

Аркадий Авраменко
(Пущинская радиоастрономическая обсерватория, Пущино)

Пламен Ангелов
(Университет Ланкастера, Великобритания)

Александр Афанасьев
(Институт проблем передачи информации РАН)

Ладиель Беллатреш
(LIAS/ISAE-ENSMA, Пуатье, Франция)

Павел Браславский
(Уральский федеральный университет, Екатеринбург)

Василий Бунаков
(STFC, Харвелл, Оксфордшир, Великобритания)

Наталья Васильева
(ЗАО «Хьюлетт-Паккард АО»)

Питер Виттенбург
(Институт психолингвистики Макса Планка, Нидерланды)

Алексей Вовченко
(ФИЦ ИУ РАН, Москва)

Владимир Голенков
(Белорусский государственный университет информатики и радиоэлектроники, Минск, Беларусь)

Владимир Головко
(Брестский государственный технический университет, Беларусь)

Евгений Гордов
(*Институт мониторинга климатических и экологических систем СО РАН, Томск)*

Ольга Горчинская
(ФОРС, Москва)

Валерия Грибова
(Институт автоматизации и управления, Дальневосточный федеральный университет, Владивосток)

Максим Губин
(Google, США)

Наталья Гулякина
(Белорусский государственный университет информатики и радиоэлектроники, Минск, Беларусь)

Юрий Демченко
(Университет Амстердама, Нидерланды)

Борис Добров
(НИВЦ МГУ, Москва)

Александр Елизаров
(Казанский федеральный университет, Казань)

Владимир Задорожный
(Университет Питтсбурга, США)

Юрий Загорулько
(Институт систем информатики им. А.П. Ершова СО РАН, Новосибирск)

Виктор Захаров
(ФИЦ ИУ РАН, Москва)

Сергей Знаменский
(Институт программных систем им. А.К. Айламазяна РАН, Ярославская обл.)

Леонид Калиниченко
(ФИЦ ИУ РАН, Москва)

Джордж Карипис
(Университет Миннесоты, Миннеаполис, США)

Надежда Киселева
(ИМЕТ РАН, Москва)

Алексей Климентов
(Национальная лаборатория Брукхавен, США)

Михаил Когаловский
(Институт проблем рынка РАН, Москва)

Владимир Кореньков
(ОИЯИ, Дубна)

Сергей Дмитриевич Кузнецов
(Институт системного программирования РАН, Москва)

Сергей Олегович Кузнецов
(НИУ ВШЭ, Москва)

Дмитрий Ландэ
(Институт проблем регистрации информации НАН Украины, Украина)

Джузеппе Лонго
(Неапольский университет, Италия)

Наталья Лукашевич
(НИВЦ МГУ, Москва)

Иван Лукович
(Университет Нови Сад, Сербия)

Олег Малков
(Институт астрономии РАН, Москва)

Яннис Манолопулос
(Университет им. Аристотеля в Салониках, Греция)

Мануэль Маццара
(Иннополис, Казань)

Ксения Найденова
(Военно-медицинская академия, Санкт-Петербург)

Дмитрий Намиот
(ВМК МГУ, Москва)

Игорь Некрестьянов
(Verizon Corporation, США)

Геннадий Ососков
(Объединённый институт ядерных исследований, Дубна)

Дмитрий Палей
(Ярославский государственный университет им. П.Г. Демидова, Ярославль)

Николай Подколодный
(ИЦиГ СО РАН, Новосибирск)

Наталия Пономарева
(Научный центр неврологии РАМН, Москва)

Алексей Позаненко
(ИКИ РАН, Москва)

Андреас Раубер
(Университет Вены, Австрия)

Тимос Селлиш
(Технологический университет Суинберна, Мельбурн, Австралия)

Владимир Серебряков
(Вычислительный центр им. А.А. Дородницына РАН, Москва)

Николай Скворцов
(ФИЦ ИУ РАН, Москва)

Владимир Смирнов
(Ярославский государственный университет им. П.Г. Демидова, Ярославль)

Валерий Соколов
(Ярославский государственный университет им. П.Г. Демидова, Ярославль)

Сергей Ступников
(ФИЦ ИУ РАН, Москва)

Александр Сычев
(Воронежский государственный университет, Воронеж)

Алексей Ушаков
(Калифорнийский университет, Санта- Барбара, США)

Александр Фазлиев
(Институт оптики атмосферы СО РАН, Томск)

Ральф Хофштадт
(Университет Билефельда, Германия)

Дмитрий Царьков
(Университет Манчестера, Великобритания, США)

Менфред Шнепс-Шнеппе
(AbavaNet, Латвия)

Георгий Чернышев
(Санкт-Петербургский государственный университет, Санкт-Петербург)

Сопредседатели программного комитета диссертационного семинара

Сергей Ступников
(ФИЦ ИУ РАН, Москва)

Сергей Герасимов
(МГУ, Москва)

Дополнительные рецензенты

Амит Гупта
(EPFL, Лозанна, Швейцария)

Владимир Иванцевич
(Университет Нови Сад, Сербия)

Дмитрий Ковалев
(ФИЦ ИУ РАН, Москва)

Борис Орехов
(ВШЭ, Москва)

Андрей Савченко
(ВШЭ, Москва)

Панайотис Смерос
(EPFL, Лозанна, Швейцария)

Иван Шанин
(ФИЦ ИУ РАН, Москва)

XIX International Conference DAMDID / RCDL'2017

DATA ANALYTICS AND MANAGEMENT IN DATA INTENSIVE DOMAINS

October 10–13, 2017, Moscow, Russia
damdid2017.frccsc.ru

COORDINATING COMMITTEE OF DAMDID/RCDL CONFERENCES

Co-Chairs

Nikolay Kolchanov,
academician RAS (Institute of Cytology and Genetics, SB RAS, Novosibirsk, Russia)

Igor Sokolov,
academician RAS (Federal Research Center "Computer Science and Control" of RAS, Russia)

Deputy chair

Leonid Kalinichenko
(Federal Research Center "Computer Science and Control" of RAS, Russia)

Members of coordinating committee

Arkady Avramenko
(Pushchino Radio Astronomy Observatory, RAS, Russia)

Pavel Braslavsky
(Ural Federal University, Yekaterinburg, Russia)

Vasily Bunakov
(Science and Technology Facilities Council, Harwell, Oxfordshire, UK)

Alexander Elizarov
(Kazan (Volga region) Federal University, Kazan, Russia)

Alexander Fazliev
(Institute of Atmospheric Optics, SB RAS, Tomsk, Russia)

Alexei Klimentov
(Brookhaven National Laboratory, USA)

Mikhail Kogalovsky
(Market Economy Institute, RAS, Moscow, Russia)

Vladimir Korenkov
(JINR, Dubna, Russia)

Mikhail Kuzminski
(Institute of Organic Chemistry, RAS, Russia)

Sergey Kuznetsov
(Institute for System Programming, RAS, Russia)

Vladimir Litvine
(Evogh Inc., California, USA)

Oleg Malkov
(Institute of Astronomy, RAS, Russia)

Archil Maysuradze
(Lomonosov Moscow State University, Russia)

Alexander Marchuk
(Institute of Informatics Systems, SB RAS, Russia)

Igor Nekrestjanov
(Verizon Corporation, USA)

Boris Novikov
(St.-Petersburg State University, Russia)

Nikolay Podkolodny
(ICaG, SB RAS, Novosibirsk, Russia)

Aleksey Pozanenko
(Space Research Institute, RAS, Russia)

Vladimir Serebryakov
(Computing Center of RAS, Russia)

Yury Smetanin
(Russian Foundation for Basic Research, Moscow, Russia)

Vladimir Smirnov
(Yaroslavl State University, Russia)

Sergey Stupnikov
(Federal Research Center "Computer Science and Control" of RAS, Russia)

Konstantin Vorontsov
(Lomonosov Moscow State University, Russia)

Viacheslav Wolfengagen
(National Research Nuclear University "MEPhI", Russia)

Victor Zakharov
(Federal Research Center "Computer Science and Control" of RAS, Russia)

General Chair of DAMDID/RCDL'2017 Conference

Igor Sokolov,
academician RAS (Federal Research Center "Computer Science and Control" of RAS, Russia)

ORGANIZING COMMITTEE

Co-chairs

Vladimir Sukhomin
(CMC MSU, Russia)

Victor Zakharov
(FRC CSC RAS, Russia)

Alexander Nevzorov
(NRNU MEPhI, Russia)

Members of Organizing Committee

Dmitry Briukhov
(FRC CSC RAS,), conference web site support

Nikolay Skvortsov
(FRC CSC RAS), CMT site support

Dmitry Namiot *(CMC MSU),*

Elena Zubareva *(CMC MSU, YelSU),*

Nikolay Skvortsov *(FRC CSC RAS),*

Alexander Elizarov *(KFU), interaction with the members of the PC and the authors in the process of collecting, reviewing, selecting and editing submissions for the conference and publishing in the local proceedings volume*

Elena Zubareva
(CMC MSU, YelSU), technical editing and publishing of the conference materials (program and the local proceedings volume)

Evgeny Morkovin
(CMC MSU), attracting participants, advertising conference

Dmitry Kovalev
(FRC CSC RAS), e-mail, contacts

Evgenia Dudareva
(FRC CSC RAS), legal and financial issues

Alexey Lifar
(NRNU MEPhI), financial support for the invited speakers from abroad

Irina Karzalova
(FRC CSC RAS), financial issues

Yulia Trusova
(FRC CSC RAS), visa support

Elena Zubareva
(CMC MSU, YelSU), the Chairperson of the local arrangements of the participants of the conference, including registration of participants, organizing lunches and coffee breaks, welcome party and the conference dinner, accommodation of the participants to stay at the MSU, the organization of excursions by the Lomonosov Moscow State University in Moscow

Alexey Lifar
(NRNU MEPhI), organization of the MEPhI choir concert for participants

Alexey Lifar
(NRNU MEPhI), organization of meetings at the airport and transfer of the conference participants

Vladimir Sukhomlin, *(CMC MSU);* Evgeny Morkovin
(CMC MSU), organization of the registration fee payment

Evgeniy Ilyushin *(CMC MSU),* Dmitry Gouriev
(CMC MSU), Vladimir Romanov *(CMC MSU), responsible for technical support, including PC, projectors, Wi-Fi at the conference rooms and at the lodging, synchronous translation support*

Alexey Lifar
(NRNU MEPhI), synchronous translation organization

PROGRAM COMMITTEE

Co-chairs

Leonid Kalinichenko
(Federal Research Center "Computer Science and Control" of RAS, Russia)

Yannis Manolopoulos
(Aristotle University of Thessaloniki, Greece)

Dmitry Namiot
(Department of Computational Mathematics and Cybernetics of the Lomonosov Moscow State University, Russia)

Members of the Program Committee

Karl Aberer
(EPFL, Switzerland)

Aleksander Afanasiev
(Institute of information transmission problems of RAS, Russia)

Plamen Angelov
(Lancaster University, UK)

Arkady Avramenko
(Pushchino Observatory, Russia)

Ladjel Bellatreche
(LIAS/ISAE-ENSMA, France)

Pavel Braslavski
(Ural Federal University, Russia)

Vasily Bunakov
(Science and Technology Facilities Council, Harwell, UK)

Georgy Chernyshev
(St.-Petersburg State University, Russia)

Yuri Demchenko
(University of Amsterdam, Netherlands)

Boris Dobrov
(Research Computing Center of Lomonosov Moscow State University, Russia)

Alexander Elizarov
(Kazan Federal University, Russia)

Alexander Fazliev
(Institute of Atmospheric Optics, SB RAS, Russia)

Vladimir Golenkov
(Belarusian State University of Informatics and Radioelectronics, Minsk, Belarus)

Vladimir Golovko
(Brest State Technical University, Belarus)

Olga Gorchinskaya
(FORS, Russia)

Evgeny Gordov
(Institute of Monitoring of Climatic and Ecological Systems SB RAS, Russia)

Valeriya Gribova
(Institute of Automation and Control Processes FEBRAS, Far Eastern Federal University, Vladivostok, Russia)

Maxim Gubin
(Google, USA)

Natalia Guliakina
(Belarusian State University of Informatics and Radioelectronics, Minsk, Belarus)

Ralf Hofestadt
(University of Bielefeld, Germany)

Leonid Kalinichenko
(FRC CSC RAS, Russia)

George Karypis
(University of Minnesota, Minneapolis, USA)

Nadezhda Kiselyova
(IMET RAS, Russia)

Alexei Klimentov
(Brookhaven National Laboratory, USA)

Mikhail Kogalovsky
(Market Economy Institute, RAS, Russia)

Vladimir Korenkov
(Joint Institute for Nuclear Research, Dubna, Russia)

Sergey Kuznetsov
(Institute for System Programming, RAS, Russia)

Sergei Kuznetsov
(HSE, Russia)

Dmitry Lande
(Institute for Information Recording, NASU, Ukraine)

Giuseppe Longo
(Naples, Italy)

Natalia Loukachevitch
(Lomonosov Moscow State University, Russia)

Ivan Lukovic
(University of Novi Sad, Serbia)

Oleg Malkov
(Institute of Astronomy, RAS, Russia)

Yannis Manolopoulos
(School of Informatics of the Aristotle University of Thessaloniki, Greece)

Manuel Mazzara
(Innopolis, Russia)

Ksenia Naidenova
(S.M. Kirov Military Medical Academy, Saint-Petersburg, Russia)

Dmitry Namiot
(Lomonosov Moscow State University, Russia)

Igor Nekrestyanov
(Verizon Corporation, USA)

Gennady Ososkov
(Joint Institute for Nuclear Research, Russia)

Dmitry Paley
(Yaroslav State University, Russia)

Nikolay Podkolodny
(Institute of Cytology and Genetics SB RAS, Russia)

Natalia Ponomareva
(Scientific Center of Neurology of RAMS, Russia)

Alexey Pozanenko
(Space Research Institute, RAS, Russia)

Andreas Rauber
(Vienna TU, Austria)

Timos Sellis Swinburne
(University of Technology, Australia)

Vladimir Serebryakov
(Computing Centre of RAS, Russia)

Nikolay Skvortsov
(FRC CSC RAS, Russia)

Vladimir Smirnov
(Yaroslavl State University, Russia)

Manfred Sneps-Sneppe
(AbavaNet, Latvia)

Valery Sokolov
(Yaroslavl State University, Russia)

Sergey Stupnikov
(FRC CSC RAS, Russia)

Alexander Sychev
(Voronezh State University, Russia)

Dmitry Tsarkov
(Independent Researcher, UK, USA)

Alexey Ushakov
(University of California, Santa Barbara, USA)

Natalia Vassilieva
(Hewlett-Packard, Russia)

Alexei Vovchenko
(FRC CSC RAS, Russia)

Peter Wittenburg
(MPI for Psycholinguistics, Netherlands)

Vladimir Zadorozhny
(University of Pittsburgh, USA)

Yury Zagorulko
(Institute of Informatics Systems, SB RAS, Russia)

Victor Zakharov
(FRC CSC RAS, Russia)

Sergey Znamensky
(Institute of Program Systems, RAS, Russia)

PhD Workshop Co-Chairs

Sergey Stupnikov
(Federal Research Center "Computer Science and Control" of RAS, Russia)

Sergey Gerasimov
(Lomonosov Moscow State University, Russia)

Additional reviewers

Andrey Savchenko
(HSE, Moscow Russia)

Amit Gupta
(EPFL, Lausanne, Switzerland)

Panayiotis Smeros
(EPFL, Lausanne, Switzerland)

Dmitriy Kovalev
(FRC CSC RAS, Moscow, Russia)

Ivan Shanin
(FRC CSC RAS, Moscow, Russia)

Vladimir Ivančević
(University of Novi Sad, Serbia)

Boris Orekhov
(HSE, Moscow, Russia)

СОДЕРЖАНИЕ / CONTENTS

Предисловие

Международная конференция «Аналитика и управление данными в областях с интенсивным использованием данных» (“Data Analytics and Management in Data Intensive Domains” (DAMDID/RCDL’2017)) проведена с 9 по 13 октября 2017 года в Московском государственном университете имени М.В. Ломоносова на факультете вычислительной математики и кибернетики.

Традиционно конференция «Аналитика и управление данными в областях с интенсивным использованием данных» представляет собой мультидисциплинарный форум исследователей и практиков из разнообразных областей деятельности людей, содействующий сотрудничеству и обмену идеями в сфере анализа и управления данными в областях исследований, движимых интенсивным использованием данных (ОИИД). На английском языке ОИИД звучит как data intensive domains (DID) в смысле 4-й парадигмы научных исследований. Подходы к анализу данных и управлению данными, развиваемые в конкретных ОИИД X-информатики (таких как X=астро, био, гео, нейро, мед, физика, химия, материаловедение и пр.), социальных наук, а также информатики, промышленности, новых технологий, финансов и бизнеса составляют предметную область конференции.

Программа конференции 2017 года, как и в предыдущих конференциях, отражает наряду с традиционной для управления данными тематикой движение в направлении науки о данных (data science) и аналитики с интенсивным использованием данных. Как и прежде, программа этого года включает специально подобранные приглашенные и ключевые доклады в быстро развивающихся ОИИД. Целью организации таких сессий также является привлечение внимания специалистов в выбранных ОИИД к конференции.

Предконференционная пленарная сессия 9 октября включает два доклада: ключевой доклад Стефано Чери (профессора Миланского политехнического института) и приглашенный доклад Золтана Саллаши (научного сотрудника Медицинской Школы Гарварда). Сессия посвящена развитию методов и средств получения генома и диагностики в различных областях (от здравоохранения до криминалистики). Стефано Чери рассматривает вопросы реализации в Европейском проекте GeCo средств секвенирования ДНК нового поколения; в докладе Золтана Саллаши обозреваются подходы к геномно-базированной диагностике в различных прикладных областях. Более подробно 10 октября Золтан Саллаши в тьюториале рассматривает применение геномной диагностики в имунотерапии рака.

Проблемы «наводнения» данными в астрономии и способы их разрешения рассматриваются в ключевом докладе Джузеппе Лонго (профессор астрофизики в Неаполитанском университете). Последний по времени представления ключевой доклад Катрин Амунтс, профессора медицины, директора Института Нейронауки и Медицины в Исследовательском Центре Juelich, посвящен большому Европейскому проекту изучения мозга человека (HBP), как флагманскому проекту по ускорению и коллаборации исследований в нейронауке.

Программный комитет конференции рассмотрел 75 заявок на конференцию и 8 заявок на диссертационный семинар. На семинар принято 5 докладов, 3 - отклонены. На конференцию 47 заявок приняты как полные статьи, 12 – как краткие, 2 – как постеры, 2 – как демо, 12 – отклонены. 59 докладов (полные и краткие) представлены в 19 сессиях, таких как проекты анализа данных в астрономии, техники Семантического Веба в ОИИД, специализированные инфраструктуры в ОИИД (1), распределенные вычисления, специализированные инфраструктуры в ОИИД (2), оценка эффективности систем, проекты анализа данных в нейронауке, специфические методы анализа данных, онтологические модели и применения (1), интеграция неоднородных баз данных, анализ гуманитарных текстов (1), проекты анализа данных в различных ОИИД, анализ гуманитарных текстов (2),

онтологические модели и применения (2), организация экспериментов в ОИИД, проекты электронных библиотек, представление и извлечение знаний, подходы к решению задач в ОИИД, применение машинного обучения. Хотя большинство докладов посвящены результатам исследований, выполняемых в исследовательских организациях, расположенных в различных местах на территории России, включая: Воронеж, Звенигород, Казань, Москву, Новосибирск, Обнинск, Омск, Орел, Переславль-Залесский, Санкт-Петербург, Томск, Ярославль, конференция становится международной. Об этом с свидетельствуют доклады (всего докладов 12 из которых 4 приглашенных), подготовленные видными зарубежными исследователями из таких стран как Армения (Ереван), Бахрейн (Манама), Беларусь (Минск), Болгария (София), Великобритания (Харвел), Германия (Дюссельдорф, Киль), Греция (Салоники), Италия (Милан, Неаполь), США (Гарвард).

Председатели Программного и Организационного комитетов конференции выражают благодарность Николаю Скворцову за обеспечение взаимодействия при помощи системы CMT с авторами присланных работ и с членами ПК – рецензентами докладов, а также за подготовку редакции сборника трудов конференции в процессе издания его печатной версии. На этом этапе проявил инициативу и оказал неоценимую помощь в редактировании текстов русскоязычных докладов профессор Федерального университета Казанского (Приволжского) федерального университета Александр М. Елизаров, которому председатели ПК выражают особую признательность. Председатели ПК также выражают благодарность членам Программного комитета за выполненную ими работу по рецензированию и отбору докладов, а также Дмитрию Брюхову за поддержку актуального содержания сайта конференции на всех этапах подготовки DAMDID/RCDL’2017.

Председатели Организационного комитета и Программного комитета конференции выражают благодарность авторам поданных на конференцию заявок, а также Российскому фонду фундаментальных исследований, Национальному исследовательскому ядерному университету МИФИ, Фонду содействия развитию интернет-медиа, ИТ-образования, человеческого потенциала «Лига интернет-медиа» за финансовую поддержку конференции.

Сопредседатели Программного комитета

Калиниченко Леонид Андреевич
(ФИЦ ИУ РАН)

Манолопулос Янис
(Университет Аристотеля, Салоники)

Сопредседатель Оргкомитета

Захаров Виктор Николаевич
(ФИЦ ИУ РАН)

Preface

In 2017 the International Conference “Data Analytics and Management in Data Intensive Domains” (DAMDID/RCDL’2017) takes place on October 9-13 in the Lomonosov Moscow State University at the department of Computational Mathematics and Cybernetics.

By tradition the “Data Analytics and Management in Data Intensive Domains” conference (DAMDID) is planned as a multidisciplinary forum of researchers and practitioners from various domains of science and research, promoting cooperation and exchange of ideas in the area of data analysis and management in domains driven by data intensive research. Approaches to data analysis and management being developed in specific data intensive domains (DID) of X-informatics (such as X = astro, bio, chemo, geo, med, neuro, physics, chemistry, material science etc.), social sciences, as well as in various branches of informatics, industry, new technologies, finance and business are expected to contribute to the conference content.

The program of the DAMDID/RCDL’2017, as it was at the previous editions of these conferences, alongside with the traditional data management topics reflects a rapid move into the direction of data science and data intensive analytics. The program of this year includes carefully selected invited keynote talks related to the fast developed DID. The respective plenary sessions are also aimed at attracting an attention of researchers in the selected DID to the conference.

Preconference plenary session on October 9 includes two talks: the keynote talk by Stefano Ceri (professor of Database Systems at the Dipartimento di Elettronica, Informazione e Bioingegneria (DEIB) of Politecnico di Milano) and the invited talk by Zoltan Szallasi (MD, a senior research scientist, the Children's Hospital Informatics Program, Harvard Medical School). The session is devoted to the development of methods and techniques of getting genoms and diagnostics in various application domains (from healthcare to criminalistics). Stefano Ceri considers the implementation issues of the new generation DNA sequencing techniques in the European project GeCo applying Big Data technologies; in the Zoltan Szallasi talk an overview of approaches to the genomic based diagnostics in various application domains is given. In more details in the tutorial given by Zoltan Szallasi on October 10 the application of genomic diagnostics in the cancer immunotherapy is presented.

The problems of data deluge in astronomy and their solution approaches are considered in the keynote talk by Giuseppe Longo (professor of astrophysics at the University of Napoli Federico II). The last but not least keynote talk is given by Katrin Amunts, MD, full professor for Brain Research and director of the C. and O. Vogt Institute for Brain Research at the Heinrich-Heine University Duesseldorf, director of the Institute of Neuroscience and Medicine (INM-1), Research Centre Juelich. This talk is devoted to the large European project HBP (Human Brain Project) as the flagship – accelerating brain science discovery and collaboration.

The conference Program Committee has reviewed 75 submissions for the conference and 8 submissions for the PhD workshop. For the workshop 5 papers were accepted and 3 were rejected. For the conference 47 submissions were accepted as full papers, 12 as short papers, 2 as posters, 2 as demos, whereas 12 submissions were rejected. According to the conference program, these 59 oral presentations (of the full and short papers) are structured into 19 sessions including Data analysis projects in astronomy, Semantic Web techniques in DID, Special-purpose DID infrastructures (two sessions), Distributed computing, System efficiency evaluation, Data analysis projects in neuroscience, Specific data analysis techniques, Ontological models and applications (two sessions), Heterogeneous database integration, Text analysis in humanities (two sessions), Data analysis projects in various DID, Organization of experiments in data intensive research, Digital library projects, Knowledge representation and discovery, Approaches for problem solving in DID, Application of machine learning. Though most of the presentations are dedicated to the results of researches conducted in the research organizations located on the territory of the Russian Federation including Kazan, Moscow, Novosibirsk, Obninsk, Omsk, Orel, Pereslavl-Zalessky, Saint Petersburg, Tomsk, Yaroslavl, Zvenigorod, the conference acquires features of

internationalization. This move is witnessed by 12 talks (four of them are invited) prepared by the notable foreign researchers from such countries as Armenia (Yerevan), Bahrain (Manama), Belarus (Minsk), Bulgaria (Sofia), Germany (Dusseldorf, Kiel), Great Britain (Harvel), Greece (Thessaloniki), Italy (Milano, Napoli), USA (Garvard).

The chairs of the Program Committee and Organizing Committee of DAMDID/RCDL'2017 express their gratitude to Nikolay Skvortsov for providing of the effective interactions by the CMT system with the authors of submissions and with the PC members reviewing the submissions, as well as for his contribution into the editing of the conference proceedings during the process of its publishing. At this stage on his own initiative a significant help in editing of the papers written in Russian was given by Alexander Elizarov, professor of the Kazan Federal University, to whom the chairs of PC express a particular thankfulness. The chairs of PC also express their gratitude to the PC members for carrying out the reviewing of the submissions and selection of the papers for presentation, as well as to Dmitry Briukhov for keeping of the up-to-date content of the conference site at all stages of the conference preparation.

The chairs of the Organizing Committee and Program Committee of DAMDID/RCDL'2017 express their gratitude to the authors of the submissions as well as to the Russian Foundation for Basic Research and the National Research Nuclear University MEPhI and the Fund "League online media" for the financial support to the Conference.

Co-chairs of the Program committee

Leonid A. Kalinichenko
(FRC CSC RAS)

Yannis Manolopoulos
(Aristotle University, Thessaloniki)

Co-chair of the Organizing committee

Victor N. Zakharov
(FRC CSC RAS)

Ключевой доклад 1

Keynote Talk 1

Data-Driven Genomic Computing: Making sense of Signals from the Genome

(Extended Abstract)

© Stefano Ceri, © Arif Canakoglu, © Abulrahman Kaitoua, © Marco Masseroli, © Pietro Pinoli

Dipartimento di Elettronica, Informazione e Bioingegneria (DEIB)
Politecnico di Milano – P.za L. Da Vinci 32,
Milano, Italy

Abstract. Genomic computing is facing a technological revolution. In this paper, we argue that the most important problem of genomic computing is tertiary data analysis, concerned with the integration of heterogeneous regions of the genome – because regions carry important signals, and the creation of new biological or clinical knowledge requires the integration of these signals into meaningful messages.

Keywords: genomic computing, high-level data models and languages, data integration.

1 Introduction

Genomics is a relatively recent science. Historically, the double helix model of DNA, due to Nobel prices James Watson and Francis Crick, was published on Nature on April 1953; and the first draft of the human genome, produced as result of the Human Genome Project, was published on Nature in February 2001, with the full sequence completed and published in April 2003. The Human Genome Project, primarily funded by the National Institutes of Health (NIH), was the result of a collective effort involving twenty universities and research centers in the United States, the United Kingdom, Japan, France, Germany, Canada, and China.

In the last 15 years, the technology for DNA sequencing has made gigantic steps. Figure 1 shows the cost of DNA sequencing in the last fifteen years; by inspecting the curve, one can note a huge drop around 2008, with the introduction of Next Generation Sequencing, a high-throughput, massively parallel technology based upon the use of image capturing. The cost of producing a complete human sequence dropped to 1000 US$ in 2015 and is expected to drop below 100 US$ in 2018.

Figure 1 Cost of DNA Sequencing, Source: NIH

Proceedings of the XIX International Conference "Data Analytics and Management in Data Intensive Domains" (DAMDID/RCDL'2017), Moscow, Russia, October 10–13, 2017

Each sequencing produces a mass of information (raw data) in the form of "short reads" of genome strings. Once stored, the raw data of a single genome reach a typical size of 200Gbyte; it is expected [1] that between 100 million and 2 billion human genomes will be sequenced by 2025, thereby generating the biggest "big data" problem for the mankind.

2 From strings to signals

Technological development also marked the generation of new methods for extracting signals from the genome, and this in turn is helping us in better understanding the information that the genome is bringing to us. Our concept of genome has evolved, from considering it as a long string of 32 billions of base pairs, encoding adenine (A), cytosine (C), guanine (G), and thymine (T), to that of a living system producing signals, to be integrated and interpreted.

The most interesting signals can be classified as:

(a) **mutations,** telling us specific positions or region of the genome where the code of an individual differs from the expected code of the "reference" human being. Mutations are associated with genetic diseases – which are inherited; mutations occur on specific positions of the genes – and other diseases such as cancer – which are also produced during the human life, and correlate with factors such as nutrition and pollution.

(b) **gene expression,** telling us in which specific conditions genes are active (i.e. they transcribe a protein) or inactive. It turns out that the same gene may have a big activity in given conditions and no activity in other.

(c) **peaks of expression,** indicating specific position of the genome where there is an increase of short reads due to a specific treatment of DNA; these in turn indicate specific biological events, such as the binding of a protein to the DNA.

These signals can be observed by using a genome browser, i.e. a viewer of the genome. All signals are aligned to a reference genome (the standard sequence characterizing human beings; such sequence is constantly improved and republished by the scientific community). The browser is open on a window of a given

length (from few bases to millions of bases), and the signals are presented as tracks on the browser; each track, in turn, show the signal – either by showing their intensity or just by showing their position. Figure 2 presents a window; the red, blue, and yellow tracks describe gene expression, peaks of expressions, and mutations. The black line indicates the position of (four) genes – these are known information, or "annotations", that can be included in the window. An interesting biological question could be: "find genes which are expressed, where there are three peaks (i.e., peaks representing three experiments are confirmed by all experiments) and with at least one mutation. Such question would, in this specific example, extract the second gene.

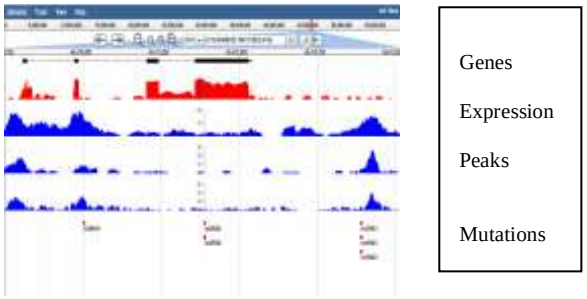


Figure 2 Signals corresponding to mutations, gene expression and peaks as shown on a genome browser

3 Tertiary Data Analysis and GeCo

Signals can be loaded on the browser only after being produced as result of long and complex bio-informatics pipelines. In particular, analysis of NGS data is classified as primary, secondary and tertiary (see Figure 3): primary analysis is essentially responsible of producing raw data; secondary analysis is responsible of extracting ("calling") the signal from raw data and align the signals to the reference genome; and tertiary analysis is responsible of a number of tasks all concerned with data integration.

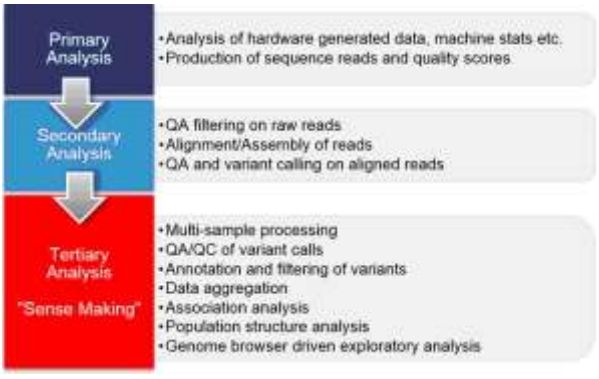


Figure 3 Classification of data analysis for genomics, http://blog.goldenhelix.com/grudy/a-hitchhiker%E2%80%99s-guide-to-next-generation-sequencing-part-2

The bioinformatics community has produced a huge number of tools for secondary analysis. So far, it has not been equally engaged in tertiary data analysis, which is clearly the most important aspect of future research.

Figure 4 shows that the number and variety of tools for secondary analysis. Instead, only four few systems are focused on tertiary analysis.

GeCo is developed by our group at Politecnico di Milano as outcome of an Advanced ERC Grant. GeCo is based on GMQL, a high-level language for genomic data management, and has a system architecture based on cloud computing, implemented on engines such as Spark and Flink [3]. **SciDB**, a scientific database produced by the spinoff company Paradigm4, supports a genomic addition focused on genomic data integration [4]. **DeepBlue**, provides easy access to datasets produced within the BluePrint consortium, with a language which is quite similar to GMQL [5]. **FireCloud,** developed by the Broad Institute, offers an integrated platform supporting cancer research [6]. All these systems already support access to a huge number of open datasets, including ENCODE and TCGA.

Figure 4 The landscape of genomic computing tools; few of them are dedicated to tertiary data analysis

This two-page abstract has set the stage for discussing why GeCo is an important project in the context of tertiary data analysis for genomics. In the full paper, we will describe some of the aspects of the GeCo project; we will focus on the GeCo API (not been presented so far). This is an important aspect of the project, as it guarantees usability of the system from multiple user and language interfaces, thereby allowing interoperability.

References

[1] Z. D. Stephens et Al.: Big Data: Astronomical or Genomical? PLoS Biol; 13(7) (2015)

[2] Kaitoua A, Pinoli P, Bertoni M, Ceri S Framework for Supporting Genomic Operations, IEEE-TC, 10.1109/TC.2016.2603980 (2016)

[3] Masseroli M, et Al.: GenoMetric Query Language: A novel approach to large-scale genomic data management. Bioinformatics 31(12), 10.1093/bioinformatics/btv048 (2015)

[4] Anonymous paper, Accelerating Bioinformatics Research with New Software gor Big Data to Knowledge (BD2K), Paradigm4 Inc. (2015)

[5] Albrecht F. et Al.: DeepBlue Epigenomic Data Server: Programmatic Data Retrieval and Analysis of the Epigenome, Nucleid Acids Research, 44/W1 (2016)

[6] https://software.broadinstitute.org/firecloud

Приглашённый доклад

Invited Talk

Development of Genomic Based Diagnostics in Various Application Domains

(Extended Abstract)

© Zoltan Szallasi

Department of Bio and Health Informatics, Technical University of Denmark,
Kemitorvet 208, 2800
Lyngby, Denmark,
Computational Health Informatics Program, Boston Children's Hospital, USA,
Harvard Medical School,
Boston, USA

zszallasi@chip.org

Abstract. We will review the revolution brought about by low cost next generation sequencing in a wide array of diagnostic and industrial applications with a special emphasis on computational requirements and big data challenges.

Keywords: next generation sequencing, big data challenges.

1 Introduction

Next generation sequencing ((NGS) has fundamentally changed modern biological research. It is, in fact, an excellent example of how gradual improvements on a powerful initial idea, Sanger's original dideoxynucleotide sequencing, can lead to such levels of quantitative increase in data production that fundamentally changes a given research field.

Virtually any nucleic acid related research question can be investigated in a comprehensive, high resolution fashion free from experimental confounding factors such as nucleic acid cross hybridization. This has produced a deluge of data on the scale of hundreds of Terabytes even for a single research laboratory. This review will survey both the various application domains of next generation sequencing and their associated computational and analytical challenges.

2 Biochemical considerations of next generation sequencing

It was recognized early on that next generation sequencing will allow querying both the genome (DNA) and the transcriptome (RNA) on a wide range of resolution. The exact sequence of nucleotides (e.g. single nucleotide polymorphisms, single nucleotide variations) and the overall architecture of the entire genome can be determined in a single experiment (one run of whole genome sequencing) [1].

A wide array of starting materials can be used for next generation sequencing. Any form of nucleic acid (DNA or RNA), from any sources (from inside the cell, from cell free biological fluids or ancient fragmented DNA) can be sequenced and quantified. Nucleic acids can be preselected as in e.g. whole exome sequencing (exon capture) or by other capture mechanisms such as specific protein beacons in ChipSeq analysis. The variations are virtually unlimited and novel approaches are constantly being added to the toolbox of biological research. This universality has led to an enormous variety of application domains.

3 Application domains of next generation sequencing

3.1 Next generation sequencing in microbiology

Since DNA is universal, next generation sequencing can be used to detect and investigate any life form from viruses to humans. This is readily exploited in the various forms of sequencing based microbial diagnostics and it has also led to a new research field, metagenomics [2]. In this, the fact that often a diverse group of microbiota live together as an "organic whole" has led to the realization that those species do not need to be isolated individually before sequencing, but the pooled DNA can be sequenced together and the sequence tags can be "sorted out" after the sequencing reaction. This ingenious idea has led to significant advances in our understanding of, for example, the microbial community of our gut flora. This method also allows monitoring sewage quality and may help to monitor and prevent disease outbreaks.

3.2 Next generation sequencing in human diseases

Genome wide association studies are a powerful method to identify germline variants associated with increased disease risk.

Remarkably, germline DNA of the fetus can be efficiently detected in maternal blood, leading to the powerful tool of prenatal testing [3]. By this, genomic, chromosomal aberrations of the fetus can be detected without virtually any risk to the mother or the fetus.

Proceedings of the XIX International Conference "Data Analytics and Management in Data Intensive Domains" (DAMDID/RCDL'2017), Moscow, Russia, October 10–13, 2017

3.3 Next generation sequencing in cancer

Cancer is a genetic disease. Accumulating mutations at various levels of the germline genome lead to malignant transformation. It is therefore, obvious, that one of the main targets of next generation sequencing is cancer diagnostics. Both germline and somatic mutations are readily identified by NGS. A great number of oncogenic mutations, many of them targetable by therapy, have thus been identified [4]. NGS data also allow us to reconstruct the evolutionary history of cancer, an issue of potentially great significance [5].
Recently, sequence analysis of liquid biopsies, essentially cell free DNA obtained from various bodily fluids, emerged as a minimally invasive tool to obtain vital information about the presence of cancer in a patient individual before or during therapy.

4 Computational and analytical challenges of next generation sequencing data

The uniform nature of the biochemical reactions and ingenuity of technical development has led to an unexpected situation. The price drop/throughput increase of next generation sequencing has significantly outpaced Moore's law over the past decade. Therefore, next generation sequencing has become more of a computational rather than a biochemical problem. The speed of data accumulation is so fast that data storage and data analysis is becoming a more and more challenging problem in modern sequencing based projects. Whole genome sequencing on a single cancer sample can easily take up hundreds of GBs of data storage. Therefore, a single study, such as the one analyzing the whole genome of 560 breast cancer cases [6], can easily produce data on at the level of hundreds of Terabytes. Such amount of data cannot possibly be downloaded for reanalysis in an efficient manner, therefore alternative solutions, such as cloud based computing had to be found.

Management of vast amounts of data is only one, mainly technical aspect of the challenges at hand.

While next generation sequencing based genomics easily qualifies as one of the main areas of big data science, in many aspects it is also markedly different from those. While in, e.g., financial data the individual variables are connected by poorly understood causative factors and in physics the entire data space is regulated by well defined, homogenous laws of physics, in biology, genomics the situation lies somewhere in between. Variables, such as genes, proteins etc. are connected by the principles of physical chemistry, but the actual parameters of those significantly vary across the various pairs of biological entities. This fact places the analysis of biological systems in the realm of robust, complex systems for which the analytical principles are poorly understood. Therefore, in order to effectively analyze the massive amounts of genomic information one needs to "front-load" the computational analysis with as much biological knowledge as possible.
We will present several strategies along those lines. In particular, we will discuss how genomics, next generation sequencing based whole genome analysis helps us to understand DNA repair pathway aberrations, and their diagnostic and therapeutic implications in cancer. We will also discuss how genomics is exploited to understand the main principles of therapeutic immune responses against cancer and how genomics, machine learning and high throughput screening are combined in an interdisciplinary environment to design effective vaccines against cancer.

5 The industrial impact of next generation sequencing

In order to satisfy the need for NGS based diagnostics a whole industry has developed during the past decade. Conferences such as the 2017 Next Generation D_x Summit, (http://www.nextgenerationdx.com) provide an excellent overview of the major trends and players.

References

[1] Mardis ER (2013) Next-generation sequencing platforms. Annu Rev Anal Chem (Palo Alto Calif) 6:287–303. doi: 10.1146/annurev-anchem-062012-092628

[2] Wooley JC, Godzik A, Friedberg I (2010) A primer on metagenomics. PLoS Comput Biol 6:e1000667. doi: 10.1371/journal.pcbi.1000667

[3] Hui L, Bianchi DW (2017) Noninvasive Prenatal DNA Testing: The Vanguard of Genomic Medicine. Annu Rev Med 68:459–472. doi: 10.1146/annurev-med-072115-033220

[4] Lawrence MS, Stojanov P, Mermel CH, et al (2014) Discovery and saturation analysis of cancer genes across 21 tumour types. Nature 505:495–501. doi: 10.1038/nature12912

[5] Jamal-Hanjani M, Wilson GA, McGranahan N, et al (2017) Tracking the Evolution of Non-Small-Cell Lung Cancer. N Engl J Med 376:2109–2121. doi: 10.1056/NEJMoa1616288

[6] Nik-Zainal S, Davies H, Staaf J, et al (2016) Landscape of somatic mutations in 560 breast cancer whole-genome sequences. Nature 534:47–54. doi: 10.1038/nature17676

Диссертационный семинар

PhD Workshop

Анализ данных

Data analysis

Volumetric Medical Image Segmentation with Deep Convolutional Neural Networks

© Manvel Avetisian © Ivan Shanin

Lomonosov Moscow State University,
Moscow, Russia

avetisian@gmail.com v08shanin@gmail.com

Abstract. This paper presents a neural network architecture for segmentation of medical images. The network trains from manually labeled images and can be used to segment various organs and anatomical structures of interest. We propose an efficient reformulation of 3D convolutions and a loss function that directly optimizes intersection-over-union metric popular in image segmentation field.

Keywords: medical image segmentation, convolutional neural networks, deep learning, convolution, loss function.

1 Introduction

Medical image is a visual representation of the interior of a body; they reveal internal anatomical structures and thus can be used for clinical analysis, intervention planning etc.

Volumetric medical images are obtained from various medical image acquisition technologies, such as computed tomography (CT), magnetic resonance tomography (MRT), etc. These images are represented by a stack of 2D image slices thus forming a 3D representation of a body [2].

Medical image segmentation is an automatic or semi-automatic process of splitting a medical image into regions, which may correspond to an organ, a tissue, a tumor, or any other anatomical structure of interest.

Some of the applications of medical image segmentation are surgical planning, virtual simulation of surgeries, tumor detection and segmentation, brain development study, functional mapping, automated classification of blood cells, mass detection in mammograms, image registration, heart segmentation and analysis of cardiac images, border detection in angiograms of coronary, etc.

Earliest medical image segmentation techniques were based on low-level processing of image data (comparing gray level values of voxels to one or multiple thresholds, edge detector filters, unsupervised clustering algorithms etc.).

Later, supervised techniques, where training data (manually labeled examples) is used to train a model, became increasingly popular. Examples of such methods are maximum likelihood and expectation maximization methods, maximum a posteriori and Markov random field methods, deformable models (active contour models, level set models), atlas-based models, conditional random field, graph cut algorithms.

Proceedings of the XIX International Conference "Data Analytics and Management in Data Intensive Domains" (DAMDID/RCDL'2017), Moscow, Russia, October 10–13, 2017

Convolutional neural networks had their applications in image segmentation, but did not gather momentum until various new techniques and computing architectures were developed. In December 2012 CNNs won ImageNet challenge for the first time. AlexNet [5] architecture proposed by Krizhevsky et al. won the competition by large margin. In subsequent years, further progress has been made [6][7]. Convolutional neural networks have become technique of choice, showing state of the art results in computer vision.

A supervised learning algorithms experience a dataset, consisting of examples, each of whichcontains features x_i and a target y_i. For example, popular Iris dataset contains measurements of various species of iris plants. A supervised learning algorithm can study the Iris dataset and learn to classify iris plants into three different species based on their measurements. In our task, $x_i$can be a computed tomography medical image, while y_i can be a segmentation of that image done by an experienced radiologist.

An artificial neural network consists of many simple units called neurons. Neurons receive and send information via weighted connections. Each neuron calculates weighted sum of inputs and applies nonlinear activation function*f* to them:

$$h(x; w, b) = f\left(\sum_i x_i * w_i + b\right).$$

Recently, one of the most popular activation functions were rectified linear units (ReLU) defined as:

$$relu(x) = \max(x, 0).$$

In a simple feed forward architecture, neurons are organized into groups called layers. Neurons in the first layer (called input layer) process information from the environment, while neurons in subsequent layers process information from previous layers. Neurons in the last layer (called output layer) produce information of interest. Because of this multi-layered structure, neural networks show very complex behavior:

$$y = h(\dots h(h(x; w_1 b_1); w_2 b_2) \dots; w_n b_n).$$

Neural networks are universal function approximators, capable of representing any function to

desired extent given enough number of neurons [1].

Convolutional neural networks (CNNs) are type of artificial neural networks specialized for processing data that has grid-like topology. Examples of such data domains include 1D time-series data or 2D or 3D images. Given two-dimensional image I and kernel K, convolution operation can S be defined as [1]:

$$(I * K)(i,j) = \sum_{m,n} I(i-m, j-n) * K(i,j).$$

A neural network computes "logit" scores, in order to convert them to probabilities, a *softmax* function is used:

$$softmax(z) = \frac{\exp(z_i)}{\sum \exp(z_i)}.$$

In order to train a neural network, we minimize a loss function $L(x, y, \hat{y}; \theta)$ with respect to θ, where $\theta = \{w_1, w_2 \dots, b_1, b_2 \dots\}$, x and y are elements of training set, and $\hat{y}$ is a prediction of the network. The loss function used in this paper will be presented in section 2.

One of the most useful (and most popular) metrics in medical image segmentation is intersection-over-union metric (IoU). For volumes A and B, IoU is defined as:

$$IoU(A,B) = \frac{|A \cap B|}{|A \cup B|}.$$

CNNs can directly classify each pixel of an image individually, given that we present to it a patch of image around pixel of interest. However, a drawback of this naïve sliding window approach is that input patches of neighboring pixels have a huge overlap, and thus some convolutions would be computed many times [2].

A significant speedup can be achieved if we present many pixels to a CNN simultaneously. One of the first implementations of this idea, that were successful in medical image segmentation, were Fully Convolutional Neural Networks (fCNN) [3]. fCNNs added upsampling layers to popular classification neural network architectures, such as AlexNet [6], VGG16 [7], and GoogLeNet [8]. This solution allowed CNN to produce a likelihood map for an entire image rather than a single pixel. The resulting neural network can be applied to an entire input volume in an efficient fashion [3].

The next iteration of fCNN idea is U-Net architecture, where a typical convolutional network architecture (contracting path) is followed by an upsampling layers (expanding path)where the size of an image is increased with upconvolutions. The resulting network forms a U-shape giving the name of the architecture. r Other major improvement are skip-connections which directly connect contracting and expanding layers. The architecture showed very good performance on different biomedical segmentation applications. It only needs a very few annotated images and has a very reasonable training time [4] due to the use of data augmentation with elastic deformations.

The 3D U-Net architecture developed ideas of U-Net further to construct a network for volumetric image segmentation that learns from sparsely annotated volumetric images. The implementation replaced all 2D convolutions of U-Net by 3D convolutions. The authors showed a successful application of the proposed method on difficult data set of the Xenopus kidney [5].

2 Method

2.1 Architecture of the neural network

The architecture of our network is based on U-Net. The input is processed by blocks of convolutional operations. The data is downscaled with maxpool operations and fed to a next convolutional block. We upscale images with upconvolutions and merged data with higher level signals before processing with another convolutional block. Figure 1summarizes the overall architecture of the neural network.

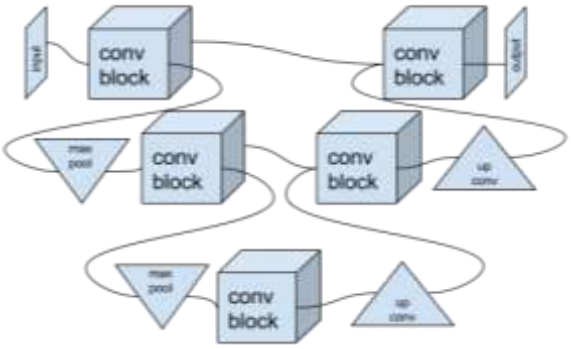


Figure 1 Architecture of proposed neural network

An convolutional block (see Figure 2) consists of four 2D convolutions along different axis. This was done to optimize processing time, as even 3x3x3 convolution has 27 parameters, while 4 3x3 convolutions have only 36 parameters. Each convolution is followed by ReLU nonlinearity and a dropout.

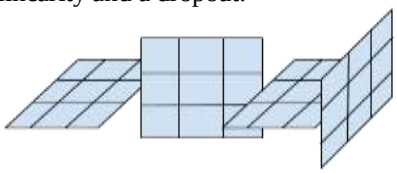


Figure 2 A convolutional block

A upconvolution layer has 1x1 kernel which upscales the data, we concatenate upscaled data with output of convoltuion block with same size.

Our experiments showed that more popular *softmax cross-entropy* function is harder to tune, as it optimizes a metric (accuracy) that we're not interested in and needs tuning of weights of examples. In our setting, IoU metric is much more informative. We extend loss function presented in [13] to multiclass setting. The loss function optimizes IoU metric directly:

$$L(x, y, \hat{y}; \theta) = \frac{\sum y_{1:} * \widehat{y_{1:}}}{\sum y_{1:} + \sum \widehat{y_{1:}} - \sum y_{1:} * \widehat{y_{1:}}},$$

where y is one-hot encoding of voxel's label, $\hat{y}$ is label probabilities outputted by the network (with *softmax* function).$y_{1:}$ denotes y without the first element.

2.2 Implementation details

The proposed method was implemented using TensorFlow library in Python 3 language [4].

A machine with Intel Core i7 6700K CPU, 32 Gb RAM, and NVidia GeForce GTX 1070 GPU was used to train a neural network and perform all experiments.

We performed an extensive search for optimal hyperparameters. Our program would select previous best hyperparameters, randomly generate new ones in interval $[0.1*p_{best}, 10*p_{best}]$, perform 5000 training steps, and select the network which showed higher IoU score on validation set. We summarised final hyperparameters that were used in Table 1.

Table 1 Best hyperparameters

dropout keep prob	0.85
l2 regularization weight	$3.0 * 10^{-4}$
learning rate	$7.6 * 10^{-5}$
channels in first conv layer	30

Adam stochastic optimization method was used. Our experiments showed that using batch normalization is not beneficial on final score [5].

3 Experiments

The Cardiac Atlas Project Dataset [9] consists of 83 volumetric MR images of heart and a mask which highlights region of interest. Figure 1 shows an example of a slice of an image from such dataset, as well a mask for that slice which highlights region of interest

Figure 3 An example of image and label from the dataset

Each image consists of 10-15 slices of various sizes, with 192x192 and 256x256 being the most frequent ones. Our model showed quality segmentation with IoU = 0.63.

4 Conclusions

Our experiment showed that convolutional neural network is capable of segmenting visually distinguishable anatomical structures on medical images. We plan to extend presented model to more medical image segmentation datasets.

Support

This research was supported by the Russian Foundation for Basic Research (grant 16-07-01028).

References

[1] Goodfellow, I., Bengio, Y., Courville, A.: Deep Learning. http://www.deeplearningbook.org

[2] Litjens, G., Kooi, T., Bejnordi, B. E.: A Survey on Deep Learning in Medical Image Analysis. https://arxiv.org/abs/1702.05747

[3] Long, J., Shelhamer, E., Darrell, T.: Fully convolutional networks for semantic segmentation (2014), arXiv:1411.4038 [cs.CV]

[4] Ronnenberger, O., Fischer, P., Brox, T.: U-Net: Convolutional Networks for Biomedical Image Segmentation. In: Medical Image Computing and Computer-Assisted Intervention -- MICCAI 2015, Part 3, pp. 234-241.

[5] Cicek, O., Abdulkadir, A., Lienkamp, S., Brox, T., Ronneberger, O.: 3D U-Net: Learning Dense Volumetric Segmentation from Sparse Annotation. In: Medical Image Computing and Computer-Assisted Intervention -- MICCAI 2015, Part 2, pp. 424-432.

[6] Krizhevsky,A., Sutskever, I.,Hinton, G. E.: Imagenetclassification with deep convolutional neural networks.In NIPS, 2012.

[7] Simonyan, K., Zisserman, A.: Very deep convolutional networks for large-scale image recognition.CoRR, abs/1409.1556, 2014.

[8] Szegedy,C., Liu,W., Jia,Y. et al.: Going deeper with convolutions.CoRR, abs/1409.4842,2014.

[9] Fonseca, C.G., Backhaus, M., Bluemke, D.A. et al.: The Cardiac Atlas Project. An imaging database for computational modeling and statistical atlases of the heart.Bioinformatics, 27(16):2288 2295, Aug 2011.

[10] Abadi, M., Agarwal, A., Barham, P.: TensorFlow: Large-scale machine learning on heterogeneous systems, 2015. Software available from tensorflow.org.

[11] Kingma, D. P., Ba, J.: Adam: A Method for Stochastic Optimization. arXiv:1412.6980 [cs.LG]

[12] Ioffe, S., Szegedy, C.: Batch Normalization: Accelerating Deep Network Training by Reducing Internal Covariate Shift. arXiv:1502.03167 [cs.LG]

[13] Rahman, A., Wang, Y.: Optimizing Intersection-Over-Union in DeepNeural Networks for Image Segmentation. In: Advances in Visual Computing – ISVC 2016 Proceedings, Part I, pp.234-244

Новый подход к определению отношения авторов коротких текстов к обсуждаемым темам на примере оценки инфляционных ожиданий

© М.Л. Андреев

Московский государственный университет имени М.В. Ломоносова,
Москва, Россия

mark.andreev@gmail.com

Аннотация. Предложен новый подход к измерению инфляционных ожиданий российского населения на основе активности населения на официальных сайтах финансово-ориентированных СМИ и их страницах в социальных сетях. Комментарии, представляющие собой короткие тексты, предварительно автоматически фильтруются на соответствие отношения к теме «инфляция» с помощью ключевых слов, составленных экспертом, и далее подвергаются анализу с помощью методов машинного обучения. Рассмотрены вычисляемые свойства отобранных комментариев, проведено тематическое моделирование с помощью методов вероятностного тематического моделирования для анализа главных тем, содержащихся в отобранных комментариях. Данный подход позволяет получать высокочастотную, экономически адекватную и обоснованную оценку инфляционных ожиданий населения РФ.

Ключевые слова: инфляционные ожидания, машинное обучение, анализ текстов, анализ естественного языка, тематическое моделирования.

A New Approach to Determining the Attitude of Authors of Short Texts to the Topics Discussed in the Texts on the Example of Estimating the Inflations Expectations

© Mark Andreev

Lomonosov Moscow State University,
Moscow, Russia

mark.andreev@gmail.com

Abstract. The paper suggests a new approach to measuring inflation expectations of the the Russian population based on its activity on official websites of financial-oriented mass media and their pages in social networks. Comments were previously automatically filtered to match the relationship to the topic "inflation" using keywords defined by the expert. Then, resulting set of comments was analyzed using machine learning methods. Simple calculated properties of the selected comments are considered; subject modeling is carried out using probabilistic thematic modeling methods to analyze the main topics contained in the selected comments. This approach makes it possible to obtain a high-frequency, economically adequate and justified estimate of inflation expectations of the Russian population.

Keywords: inflation expectations, machine learning, text analysis, natural language analysis, thematic modeling.

1 Введение

Под короткими текстами будем понимать комментарии пользователей в социальных сетях и на страницах официальных ресурсах СМИ. Короткие тексты характеризуется малым числом тем. Каждый комментарий имеет автора, время публикации и ссылку на статью, к которой он был оставлен. В зависимости от источника автор может иметь не только имя, но и другие атрибуты, например, место проживания, информацию о социальном графе и информацию о сообществах, в которых он состоит.

Под темой будем понимать совокупность слов, образующих смысловую повестку. При этом одна большая тема может включать в себя меньшие. Для выделения тем предлагается использовать два

Труды XIX Международной конференции «Аналитика и управление данными в областях с интенсивным использованием данных» (DAMDID/ RCDL'2017), Москва, Россия, 10–13 октября 2017 года

подхода: поиск темы по ключевым словам и с использованием тематического моделирования.

Выделение темы по ключевым словам требует формирования регулярного выражения. Такой подход ограничен списком тем, которые описал исследователь. Его результатом является список комментариев, относящихся к заданной теме. Для дальнейшего исследования полученных комментариев предлагается использовать тематическое моделирование.

Тематическое моделирование позволяет выделить из текстов заданное заранее количество тем. Уменьшая количество тем, исследователь будет получать более глобальные темы, увеличивая их количество, будет получать подтемы больших тем [4].

Под отношением к теме будем понимать как факт упоминания данной темы пользователем, так и эмоциональную окраску сообщений, в которых упоминается тема.

Оценка тональности комментариев возможна как по средствам заранее сформированных словарей, содержащих информацию о тональности каждой словоформы [7], так и с помощью методов машинного обучения, требующих размеченных текстов на предмет их тональности [6]. В данной работе рассматривается второй подход, не требующий кропотливой работы лингвиста.

Таким образом, предлагается фильтровать исходную совокупность комментариев на отношение к исследуемой теме, представленной в виде регулярного выражения, а затем исследовать статистики вычислимых свойств комментариев: количество упоминаний темы в единицу времени, количество эмоционально окрашенных комментариев в единицу времени. Для исследования тем, являющихся частью исследуемой, предлагается использовать тематическое моделирование.

Для иллюстрации работы предложенного подхода рассмотрим задачу измерения инфляционных ожиданий населения РФ на основе его активности на официальных сайтах финансово-ориентиро-ванных СМИ и их страницах в социальных сетях.

В экономике инфляционными ожиданиями называют предполагаемые уровни инфляции, основываясь на которых производители и покупатели строят свою будущую ценовую и кредитно-финан-совую политику [1]. Влияние на инфляционные ожидания оказывает ЦБ РФ в рамках режима инфляционного таргетирования. Инструментом воздействия с сентября 2013 г. является ключевая ставка. Кроме того, среди косвенных инструментов воздействия на инфляционные ожидания ЦБ РФ использует информационную политику, постоянно объясняя населению свои действия и дальнейшие планы. Традиционный метод оценивания инфляционных ожиданий подразумевает проведение опросов. Используя данный подход, агентство ООО «ИНФОМ» оценивает инфляционные ожидания населения РФ. Главные недостатки такого подхода состоят в низкой частоте обновления индекса (раз в месяц); ограниченности выборки, которая состоит всего из 2000 домохозяйств; скорости публикации индекса (результаты опросов ООО «ИНФОМ» запаздывают примерно на две недели после проведения опросов за счет необходимости обработки полученных данных); отсутствие возможности пересчета показателей при изменении методологии опросов высоких издержках при построении.

Подход, предлагаемый в данной статье, лишен вышеупомянутых недостатков. Высокая частота обновления индекса обеспечивается автоматизацией сбора данных и последующим анализом на высокопроизводительном вычислительном кластере. Выборка ограничена только количеством пользователей, комментирующих новости в отобранных источниках. Сохранение потока сообщений в специальное хранилище позволяет обновить индекс при изменении методологии, без повторного сбора данных.

Чувствительность индекса к изменениям инфляционных ожиданий населения отчетливо видна на графике его изменения – по ситуации в конце 2014 – начале 2015 годов. Этот период характеризуется всплеском волнения населения. Экономическое обоснование полученного индикатора детально рассмотрено в [1].

2 Подход к построению системы

2.1 Концептуальное описание

Весь процесс построения индекса инфляционных ожиданий населения можно разделить на два логических этапа: сбора данных и анализа полученных данных. Общим для двух этапов может быть хранилище, в которое поступают данные из модуля сбора данных, а затем обрабатываются модулем анализа данных. Результат работы модуля анализа данных сохраняется в хранилище отчетов.

Основываясь на данной концепции, рассмотрим два варианта организации такой системы. В первом случае система будет сохранять собранные данные в базу данных, во втором – отправлять данные в очередь, из которой модуль анализа данных будет забирать сообщения.

Для реализации концептуального прототипа был выбран первый вариант, подразумевающий последовательную работу модулей: вначале собираются все данные, потом анализируются. Данный подход имеет более простую реализацию, чем второй, а также позволяет оперировать сразу со всей выборкой данных.

Второй подход, основанный на очереди сообщений, подразумевает батчевую обработку данных на лету. Такой подход более производителен: большая скорость обработки данных и отсутствие требования хранения всех данных в оперативной памяти.

Первый метод будем называть офлайновым,

второй – онлайновым, исходя из скорости обработки поступающих данных.

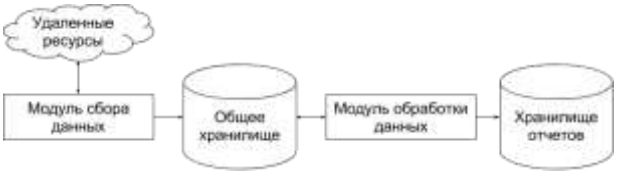


Рисунок 1 Концептуальная схема системы обработки данных

2.2 Система с офлайн обработкой данных

Рассмотрим подробнее первый способ организации приложения. Модуль сбора данных состоит из «краулера», собирающего содержимое интернет-ресурсов и очереди задач, в которую «краулер» помещает ссылки на станицы, которые он планирует посетить в дальнейшем. Загруженные страницы программа сбора сохраняет в формате json в NoSQL СУБД MongoDB. Использование данной базы продиктовано необходимостью хранить структуры данных, содержащие вложенные поля и имеющие непостоянную структуру.

Особенностью данного модуля является поддержка режима распределенного сбора данных за счет наличия внешней очереди задач, реализованной с помощью сервера очередей RabbitMQ.

Для реализации «краулера» использовался язык Java, с помощью которого было построено приложение, использующее многопоточные возможности языка для ускорения сбора данных. Отказ от выбора готового решения был обусловлен необходимостью собирать данные из неоднородных источников, что требует персонального подхода к извлечению данных.

Рисунок 2 Схема реализации системы с офлайн обработкой данных

Модуль анализа данных был написан на языке Python с использованием библиотек анализа данных Pandas, Matplotlib, Scikit-learn. Исходный код исследований хранился в формате блокнотов Jupyter. Для изоляции окружения использовался Docker, такой подход позволил добиться воспроизводимости результатов, несмотря на возможные обновления рабочей системы, влияющие на реализацию алгоритмов машинного обучения.

Для вторичного анализа данных, полученный от основного модуля обработки данных, использовался язык R, исходные коды которого так же хранились в формате блокнотов Jupyter. Использование данного инструмента вызвано его популярностью в среде аналитиков, как следствие большего разнообразия модулей для статистического анализа, чем в экосистеме Python.

2.3 Система с онлайн обработкой данных

Рассмотренная ранее система наиболее оптимальна для построения исследовательского прототипа, призванного проверить базовую гипотезу. Однако для использования системы в условиях высокой нагрузки – большого потока данных от множества «краулеров» – данная система плохо пригодна. По этой причине предлагается рассмотреть потоковую обработку данных, подразумевающую отправку данных в очередь, а не напрямую в базу данных. Из очереди сообщений должны формироваться «батчи» данных, которые следует отправлять в систему распределенной обработки данных, например, Apache Spark [3]. Одновременно с этим стоит сохранять данные в специальные долгосрочные хранилища, имеющие пониженную цену на хранение данных по сравнению со стандартными облачными хранилищами. Примером долгосрочного хранилища является Amazon Glacier, Azure LRS. Архивирование данных позволяет воспроизвести вычисления, полученные ранее.

Рисунок 3 Схема реализации системы с онлайн обработкой данных

3 Построение индекса инфляционных ожиданий

Для построения индекса инфляционных ожиданий использовались не все комментарии, а лишь относящиеся к теме «инфляция». Для фильтрации целевых комментариев использовались регулярные выражения, составленные экспертом-экономистом. В анализе оценивалось как абсолютное число целевых комментариев в единицу времени, так и их вычисляемые свойства: эмоциональный окрас, тематика. Тональность комментария можно рассматривать как отношение пользователя к проблеме, рассматриваемой в статье, к которой был оставлен комментарий. Абсолютное число эмоционально окрашенных и нейтральных комментариев в единицу времени оказалось коррелированным с индексом, предоставляемым ООО «ИНФОМ».

3.1 Оценка тональности комментариев

Для оценки тональности комментариев использовались методы машинного обучения, в

частности логистическая регрессия, метод опорных векторов. Для обучения классификаторов использовалась размеченная выборка, состоящая из русскоязычных сообщений твиттера [2]. Классификатор решал задачу бинарной классификации разделения комментариев на классы «негативный» и «позитивный». На основе вероятности отношения к классу «негативный», предоставляемой обученным классификатором, принималось решение об отнесении комментариев к трем классам «позитивный», «нейтральный», «негативный». Дискретизация проводилась на основе принадлежности к полуинтервалам и отрезку: [0, 0.25) [0.25, 0.75] (0.75, 1].

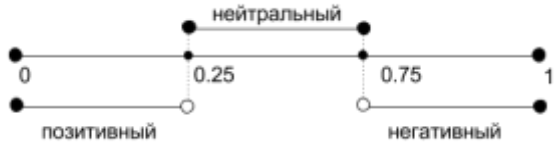


Рисунок 4 Дискретизация вероятности отнесения к классу «негативный комментарий»

Ниже представлены результаты оценки качества классификации различных моделей и методов предобработки данных, вычисленные на основе выборки сообщений из русскоязычного твиттера [2] с помощью метода перекрестной проверки с разбиением на 10 частей.

Таблица 1 Оценка качества моделей

Алгоритм	Предобработка	Верность,%
Лог регрессия	Мешок слов	76.7
	TF-IDF	76.0
Метод опорных векторов	Мешок слов	75.0
	TF-IDF	76.3

Полученная модель, вычисляющая количество эмоционально окрашенных комментариев, была противопоставлена результатам индикатора на основе опросов (традиционных подход).

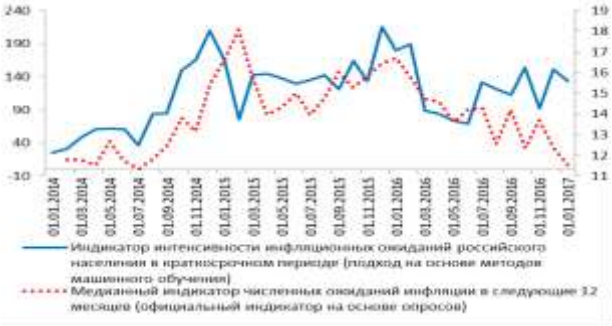


Рисунок 5 Классический индикатор и индикатор на основе методов машинного обучения

Из Рис. 4 видно, что полученный индикатор опережает индикатор, полученный традиционным путем.

3.2 Тематическое моделирование комментариев

Для анализа содержания комментариев было предложено использовать вероятностное тематическое моделирование. Такой подход позволил нам рассмотреть темы и их долю в различные периоды времени, избежав чтения всех комментариев экспертом. Вместо этого предстоит лишь рассмотреть небольшое число тем. Каждая тема представлена ключевыми словами, характеризующими ее. Для использования математических моделей тематического моделирования требуется предварительно обработать текст: нормализовать слова и удалить стоп слова и спец. символы. Для решения данной задачи использовалась библиотека NLTK и pymorphy2.

В качестве инструмента для вероятностного тематического моделирования был использован BigARTM [4], который реализует модуль ARTM (Аддитивная регуляризация тематических моделей) в качестве математической модели взаимодействия документов, терминов и тем.

Классические модели тематического моделирования малоинтерпретируемы на коротких текстах. По этой причине использовалось нестандартное представление документов – WNTM [5], которое рассматривает взаимную встречаемость слов: для каждого слова рассматривается его локальный контекст.

Количество тем для построения модели, настраивается пользователем самостоятельно. При этом увеличение числа тем ведет к слиянию менее популярных тем в одну.

На данный момент авторами не выработана окончательная методика визуализации тем с привязкой ко времени. Концептуальное видение решение данной проблемы изображено на Рис. 6. График отражает общий интерес к теме, выражающийся в абсолютном количестве сообщений, содержащих паттерн темы (удовлетворяют регулярному выражению). На оси абсцисс отложены даты, на оси ординат – абсолютное количество сообщений. Окружность, расположенная в один из моментов времени, демонстрирует момент времени, в окрестности которого производится детализация подтем посредством тематического моделирования. Список тем, характеризующийся ключевыми словами, изображен в рамке под графиком.

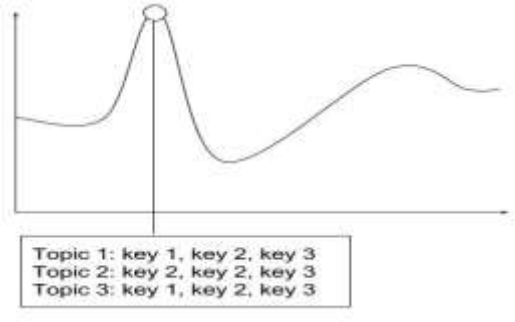


Рисунок 6 Концепция визуализации тематической модели

Авторы также планируют разработать подход для визуализации тем с учетом географии пользователей, оставляющих комментарии.

4 Заключение

Разработан подход для получения высокочастотной оценки инфляционных ожиданий населения РФ, который был реализован в виде прототипа системы анализа комментариев. Полученные практические результаты теоретически обоснованы и опубликованы в тематическом журнале [1].

Для решения данной задачи был самостоятельно реализован модуль сбора данных, имеющий специализированные компоненты извлечения данных для конкретных источников данных. Особенностью этого модуля является возможность организации распределенного сбора информации. Исследовательский код, расположенный в модуле анализа данных, использовал готовые библиотеки анализа данных. Авторы видят возможность реализации собственной тематической модели на базе BigARTM, которая бы учитывала дополнительные факторы, например, географию пользователей.

Представленный подход к построению индикатора позволяет расширять список источников комментариев, а также внести изменения в модуль анализа данных уже в процессе эксплуатации системы, уточняя значение индекса посредствам повторного вычисления с использованием новых данных.

Область применения данного подхода может выходить за рамки оценки инфляционных ожиданий и использоваться для отслеживания интереса пользователей к различным темам. Примером таких тем может быть отношение пользователей к коммерческим организациям или публичным лицам. Отслеживание интереса позволит оперативно информировать PR агентства об имидже клиента.

Подход к построению системы, представленный в статье, позволяет адаптировать систему для работы с большими данными. Увеличение охвата пользователей приведет к более точной оценке отношения пользователей к теме.

Авторами не решена проблема визуализации тематической модели для комментариев с учетом времени их публикации. Планируется модифицировать метод фильтрации комментариев, который бы учитывал место проживание пользователя. Отдельной подзадачей является идентификация ботов среди комментаторов для исключения их из рассмотрения, либо выделения в отдельную группу для информирования исследователя об их наличии.

Поддержка

Работа выполнена при поддержке РФФИ (грант 16-07-01028).

Литература

[1] Голощапова, И., Андреев, М.: Оценка инфляционных ожиданий российского населения методами машинного обучения. Вопросы экономики, (6), сс. 71-93. Некоммерческое партнерство «Редакция журнала «Вопросы экономики»» (2017)

[2] Рубцова, Ю. Построение корпуса текстов для настройки тонового классификатора. Программные продукты и системы, (1), сс. 72-78. Научно-исследовательский институт «Центрпрограммсистем» (2015)

[3] Клеменков, П. Пайплайн машинного обучения на Apache Spark. Конференция HighLoad++ (2016)

[4] Vorontsov, K. BigARTM: Open Source Library for Regularized Multimodal Topic Modeling of Large Collections. Int. Conf. on Analysis of Images, Social Networks and Texts, pp. 370-381. Springer International Publishing (2015)

[5] Zuo, Y., Jichang Z., Ke X. Word Network Topic Model: a Simple But General Solution for Short and Imbalanced Texts. arXiv preprint arXiv:1412.5404 (2014)

[6] Pang, B., Lee, L., Vaithyanathan, S.: Thumbs up?: Sentiment Classification using Machine Learning Techniques. Proc. of the ACL-02 Conf. on Empirical Methods in Natural Language Processing, 10, pp. 79-86. Association for Computational Linguistics (2002)

[7] Rao,. Y.: Building Emotional Dictionary for Sentiment Analysis of Online News. World Wide, (4), pp. 723 (2014)

[8] Chang, J. Reading tea leaves: How Humans Interpret Topic Models. Advances in Neural Information Processing Systems, pp. 288-296 (2009)

Автоматическое выделение признаков в задаче классификации сигналов

© В.А. Викулин

Московский государственный университет имени М.В. Ломоносова,
Москва, Россия

va.vikulin@physics.msu.ru

Аннотация. Рассмотрена задача классификации сигналов. Предложен стохастический алгоритм, позволяющий выделять качественное признаковое пространство в этой задаче. Показан принцип разложения признаков сигнала через базисные функции и представлен примерный набор базисных функций, который можно использовать в задачах анализа сигналов. Алгоритм строит каждый признак с помощью максимизации некоторого функционала качества, оптимизируя данное разложение. Предложено несколько вариантов таких функционалов качества. Стохастически проводя такую процедуру, можно синтезировать качественное признаковое пространство. Алгоритм проверен на задаче классификации ЭКГ сигналов, в которой по кардиограмме пациента определялось наличие у него ишемической болезни сердца.

Ключевые слова: анализ сигналов, выделение признаков, задача классификации.

Automatic Feature Extraction for Signals Classification

© V. Vikulin

Lomonosov Moscow State University,
Moscow, Russia

va.vikulin@physics.msu.ru

Abstract. The article is concerned with the signal classification problem. The article suggests an algorithm which allows to create feature space in this task. It shows a strategy performing features with the basic functions, some set of basic functions provided to be used in multiple signal processing problems. The algorithm creates each feature by maximizing some function of feature quality and some sets of possible feature quality metrics for solving signal processing tasks are recommended. If you repeat this procedure randomly you will create feature subspace. The algorithm was tested on ECG classification problem in which algorithm defined the presence of coronary disease in patients.

Keywords: signal processing, feature extraction, classification.

1 Введение

Сигнал – последовательность измерений некоторой величины. Задача классификации сигналов часто встречается во множестве различных прикладных задач – от медицины до приборостроения [5].

Не так давно на рынке стало доступно множество мобильных приборов, которые могут непрерывно записывать кардиограмму человека. Количество подобных приборов растет с каждым днем, ведь такие медицинские измерения не только удобны, но и позволяют своевременно обнаружить болезнь. Необходимо не только быстро обрабатывать все эти показания, но и быстро находить среди них больных людей, а это возможно сделать только методами анализа сигналов. Разработка алгоритмов классификации сигналов становится важной и актуальной задачей. Одним из популярных подходов к решению задачи классификации сигналов является нахождение оптимального признакового пространства, в котором объекты (сигналы) могут наиболее просто быть разделены с помощью классических алгоритмов классификации.

Рассмотрим постановку задачи классификации. Пусть заданы множество объектов X, множество допустимых ответов Y, и существует функция $y: X \to Y$, значения которой известны только на конечном подмножестве объектов $\{x_1 \dots x_l\} \subset X$. Задача заключается в том, чтобы по имеющимся парам объект–ответ восстановить исходную зависимость, то есть построить решающую функцию $a: X \to Y$, которая приближала бы целевую

Труды XIX Международной конференции «Аналитика и управление данными в областях с интенсивным использованием данных» (DAMDID/ RCDL'2017), Москва, Россия, 10-13 октября 2017 года

функцию y, причём не только на известных объектах, но и на всём множестве X. Признаком объекта x назовем результат измерения характеристики объекта. Другими словами, признаком называется отображение $f : X \to D_f$, где D_f – множество допустимых значений признака. Нахождению оптимального множества таких отображений $\{f\}$ посвящена настоящая работа.

В данном случае под множеством объектов X будем всегда иметь в виду множество сигналов, то есть конечную последовательность вещественных чисел, а под множеством допустимых ответов Y – двухэлементное множество $\{-1, 1\}$.

Нахождение оптимального признакового пространства является крайне сложной задачей. В большинстве случаев в задачах анализа сигналов для конструирования этого пространства применяются классические приемы, основанные в своем большинстве на преобразовании Фурье и вейвлет-преобразовании [6, 8]. Такой подход имеет ряд существенных недостатков. Он требует глубокого понимания от исследователя природы сигнала, и исследователь должен сам подбирать необходимое спектральное разложение; не существует некоторого универсального преобразования, которое бы позволяло всегда выделять оптимальное признаковое пространство из сигнала. Из-за этого недостатка те качественные признаки, которые были найдены в предыдущей задаче анализа сигнала, в новой задаче могут быть абсолютно неприменимы. В следующей задаче анализа сигналов исследователю необходимо с нуля конструировать признаковое пространство, опираясь исключительно на свою интуицию и опыт, полученный при решении предыдущих задач

Данная работа посвящена методу автоматического построения признакового пространства с помощью максимизации критерия качества признака (здесь и далее признаком сигнала будем называть любую вещественную функцию от сигнала). Нами использовался метод оптимизации, который является обобщением «жадного поиска», но использование конкретно этого метода оптимизации совершенно не обязательно. Поиск оптимального признакового пространства при этом являлся стохастическим, то есть в самом алгоритме заложена рандомизация, что позволяет постоянно генерировать новые признаки, отличающиеся от предыдущих. Критериев качества для оценки признака было несколько, и они тоже выбирались для каждого признака случайно. Это также помогало генерировать непохожие друг на друга признаки, так как не существует универсального метода оценки качества, при этом нельзя максимизировать их все сразу. Благодаря стохастике, данный алгоритм позволял за N итераций почти всегда найти N непохожих друг на друга признаков. Далее синтезированное множество признаков может использоваться любым классическим классификатором.

Данный подход уже применялся несколько раз в анализе сигналов [1, 3, 4, 7]. В этих работах используется генетический алгоритм для нахождения оптимального признакового пространства. Генетический алгоритм является алгоритмом оптимизации, который базируется на механизмах, в какой-то степени аналогичных механизмам эволюции в живой природе. В качестве функции, которая оптимизируется генетическим алгоритмом, выступает какая-либо мера качества признака. Например, качество предсказания алгоритма, построенного на синтезированном признаке на кросс-валидации. Основным недостатком данных работ является четкая привязка как к методу оптимизации, так и к выбору оценки качества признака. Из-за жесткой привязки к оценке качества генетический алгоритм является хорошим выбором, потому что он не старается наивным образом подобрать себе лучшее решение, как это делает, например, жадный алгоритм. Если бы в этих работах использовался «жадный» алгоритм, то признаковое пространство было бы бедным и все время одинаковым, так как этот алгоритм не обладает нужной вариативностью. Одной из важнейших задач данной работы является построение метода, в который легко бы встраивался абсолютно любой метод оптимизации, то есть предлагается обеспечивать нужную вариативность не с помощью метода оптимизации, а с помощью стохастической природы поиска оптимального признакового пространства. В этом случае метод оптимизации может быть любым, он не будет определяющим в конструкции.

2 Стохастический алгоритм синтеза признакового пространства

Рассмотрим предлагаемый алгоритм синтеза признакового пространства.

2.1 Представление признака через базисные функции

Напомним, что признаком сигнала называется функция от сигнала, которая ставит в соответствие сигналу какое-то число. Будем раскладывать каждую такую функцию через набор заранее определенных базисных функций, в рамках которых мы и будем проводить оптимизацию. Таким образом, выбор базисных функций однозначно определит пространство, в котором будет происходить оптимизация. Каждый признак при этом будет представлять собой суперпозицию базисных функций, которые будут применяться поочередно, формируя в итоге значение признака. Пусть мы выбрали множество $\{b\}$ мощности N базисных функций, тогда любой признак сигнала может быть представлен в виде:

$$f(x) = [b_1][b_2] \dots [b_{last}](x), \quad (1)$$

где b_i – очередная базисная функция из множества $\{b\}$, прямоугольные скобки используются для

разделения базисных функций и отдельного смысла не несут. Далее будем подразумевать, что функции в выражении (1) применяются слева направо. Эта форма записи не согласуется с привычными правилами записи подобных выражений в математике, но была выбрана из соображений наглядности.

Заметим, что в формуле (1) каждый признак может быть представлен через любое число базисных функций. Конкретно взятая базисная функция может быть использована в представлении признака неограниченное, но обязательно конечное число раз.

Отметим также, что если мы определили, что признаком сигнала является число, то последняя из базисных функций в формуле (1) обязана быть вида $b_{last}: R^m \to R$. Остальные функции должны быть согласованы по областям задания и областям значений: $b_i: R^{m_i} \to R^{m_{i+1}}, b_{i+1}: R^{m_{i+1}} \to R^{m_{i+2}}$. В данной работе из-за специфичности задачи анализа сигналов любой признак описывается не формулой (1), а ее несколько усложненным вариантом, что позволяет как сузить оптимизируемое пространство, так и использовать априорные знания о том, какие базисные функции вообще должны применяться в задаче обработки сигналов, в каком порядке они должны применяться.

- Функции инициализации – множество $\{i\}$. Это функции, с которых должен начинаться каждый признак в представлении (1). В представлении сигнала должна быть ровно одна функция инициализации. В множество функций инициализаций стоит включить те преобразования, которые в предметной области чаще всего используются для предобработки данных, это позволит напрямую использовать знания о предметной области при поиске признакового пространства. В задачах анализа сигналов часто сначала делают предварительную обработку сигналов: сглаживание или применение фильтра низких частот. Это функции $i: R^m \to R^n$.
- Функции трансформации – множество $\{t\}$. Это функции, которые отвечают за преобразования сигнала, который прошел через инициализацию. В каждом сигнале их может быть любое количество, число функций трансформаций может быть ограничено только из соображений вычислительной сложности получаемых признаков. Эти функции представляют собой по большей части нелинейные преобразования. Это функции $t: R^m \to R^n$
- Функции агрегации – множество {a}. Для того чтобы получить из сигнала число, необходимо в конце цепочки базисных функций поставить функцию, которая бы агрегировала всю полученную информацию в одно число, поэтому необходимо ввести функции агрегации. Функциями агрегации могут быть, например, среднее значение последовательности, максимальное значение последовательности и так далее. Это функции $a: R^m \to R^n$.

Таким образом, формула (1) может быть переписана в виде

$$f(x) = [i][t_1][t_2] \dots [t_{last-2}][a](x), \quad (2)$$

где i – функция инициализации t_j – какая-то из функций трансформации, a – функция агрегации.

Задача генерации признакового пространства заключается в том, чтобы найти X признаков, представимых в форме (2), которые были бы оптимальны с точки зрения оценки качества алгоритма, обученного на этом пространстве признаков. В свою очередь это означает, что для каждого из N признаков необходимо найти функцию инициализации, последовательность функций трансформации и функцию агрегации для данного признака.

Примеры функций инициализации: тождественная; медианное сглаживание; фильтр верхних частот; фильтр нижних частот.

Примеры функций трансформации: логарифмирование; возведение в степень; конечная разность; абсолютное значение; стандартизация сигнала (вычитание среднего и деление на дисперсию).

Примеры функций агрегации: среднее значение; медиана; дисперсия; максимум, минимум. центр масс сигнала (скалярное произведение индексов на значения сигнала).

2.2 Оценка качества признака

Чтобы построить хорошее признаковое пространство, удовлетворяющее условию (2), необходимо четко определить критерий, по которому будет проходить поиск нового признака. Таким образом, нам необходимо ввести критерий качества признака. Такие критерии сильно связаны с методами фильтрации признаков, которых на данных момент известно уже немало. Похожие подходы можно использовать и в оценке качества признака.

Самый простой способ оценить качество признака – проверить, насколько статистически признак связан с целевой переменной. В этой области существует невероятно большое число исследований. Перечислим лишь несколько способов, которые в дальнейшем будем использовать для экспериментов.

- Количество неправильно ранжированных пар целевой переменной при сортировке ее по значениям данного признака. Самая простая оценка качества. Полагаем, что значения признака есть выход некоторого классификатора. Отсортируем целевую переменную по данному признаку и проверим качество этой сортировки.

- Корреляция Пирсона между целевой переменной и признаком, то есть мера линейной зависимости признака от целевой переменной. При этом разумно брать модуль, так как нам не важен знак этой линейной зависимости.
- Взаимная информация между целевой переменной и признаком, то есть величина, описывающая количество информации, содержащегося в целевой переменной относительно признака. В качестве оценки качества разумно брать нормированное на отрезок [0,1] значение.

Статистические методы обладают важным достоинством – они очень быстро считаются. Из-за этого они получили широкое распространение в задачах, где признаковое пространство состоит из огромного числа признаков, но при этом не очень важно выделить оптимальное подмножество признаков из пространства, а гораздо важнее убрать совершенно бесполезные или даже вредные признаки. Основным недостатком этих методов является недостаточная описательная способность, любой статистический критерий не способен исчерпывающе описать степень зависимости одной величины от другой, очень высок риск ошибки в оценке качества.

Существует другой обширный класс методов оценки качества признаков, который проверяет качество алгоритма, обученного на одном этом признаке. В экспериментах использовался один из самых простых алгоритмов – алгоритм *k* ближайших соседей (*k* nearest neighbors, KNN). Описать этот алгоритм довольно просто – объект относится к тому классу, к которому относится большинство из его *k* соседей, то есть *k* ближайших к нему объектов обучающей выборки, в данной работе использовалась стандартное евклидово расстояние. Оценка качества проводилась методом скользящего контроля с исключением объектов по одному (leave-one-out, LOO). Это очень популярный метод оценки качества алгоритма *k* ближайших соседей. В этом методе каждый объект по очереди исключается из обучающей выборки, для него происходит предсказание, вычисляется оценка качества, а затем это качество усредняется.

Последний метод заключается в проверке того, что признак хорошо используется алгоритмом. В экспериментах использовался алгоритм дерева решений. К тестируемому признаку прибавлялся случайный признак, затем на этих двух признаках строилось дерево решений фиксированной глубины, оценивалось, во сколько раз тестируемый признак лучше, чем случайный признак, с помощью оценки уменьшение impurity (impurity – мера качества сплита, которая вычисляется при выборе разбиения в дереве) по разбиениям дерева решений.

2.3 Схема метода

Алгоритм (1) описывает работу метода с помощью псевдокода. Метод работает так:

- Случайно взяли *k*-элементную подвыборку из множества сигналов.
- Применили к этой подвыборке случайную функцию инициализации.
- Выбрали случайный критерий качества признака.
- Установили параметры в функциях трансформации. Например, возведение в степень *p* имеет параметр *p*. В экспериментах параметры трансформации брались случайно из заранее выбранного множества, но можно использовать любой другой подход.
- Нашли новый признак методом оптимизации. Если нужно больше признаков, начали процесс сначала (с новой случайной подвыборки).

Алгоритм 1 Стохастический алгоритм синтеза признакового пространства

```
function find_features(sigs, N, k, init_funcs,
trans_funcs, agg_funcs, criteria)
  i = 0
  features = {}
  while i !=N  do:  // Ищем N признаков
    subs = random_subsample(sigs, k)
    new_init = get_random(init_funcs)
    init_subs = new_init(subs)
    new_crit = get_random(criteria)
    set_parameters(trans_funcs)
    new_feat = optimize(init_subs, new_crit, \
                        trans_funcs, agg_funcs)
    if new_feat not in features then:
      i = i + 1
      features.insert(new_feat)
  return features
```

Построение признака по случайной подвыборке решает одновременно несколько задач. Признаки будут не очень похожи друг на друга, так как они подстраивались под разные множества. Уменьшается риск построить признаковое пространство, которое работает только на определенном наборе объектов. Позволяет избавиться от проблемы несбалансированных классов, можно брать подвыборку с равным количеством объектов каждого класса. Уменьшает вычислительную сложность оценки качества признака, которая во многих методах очень большая.

Алгоритм (2) иллюстрирует работу жадного оптимизатора. Он наращивает функции трансформации жадным образом. Наращивает до тех пор, пока не превысит заранее установленный лимит, или пока качество не перестанет расти. При добавлении новой трансформации просматриваются все возможные функции агрегации.

Алгоритм 2 «Жадный» оптимизатор

```
Function optimize(init_subs, new_crit,
trans_funcs, agg_funcs)
  best_qual = -inf
  found_trans = {}
```

```
features = {}
while len(found_trans) != MAX_SIZE do
  for new_trans in trans_funcs do
    found_better = False
    for new_agg in agg_funcs do
      feature = create(found_trans + \
                {new_trans}, new_agg)
      if qual(feature) > best_qual then
        found_better = True
        new_best_trans = new_trans
        best_agg = new_agg
  if not found_better then
    break
  found_trans.insert(new_best_trans)
return found_trans, best_agg
```

3 Вычислительные эксперименты

Эксперименты проводились на сигналах, которые представляют собой электрокардиограммы пациентов. Для каждой кардиограммы известно, болен ли пациент ишемической болезнью сердца. Это классическая задача бинарной классификации, где класс 1 означает, что пациент с данной кардиограммой болен, класс -1 – здоров. Выборка состояла из 1798 сигналов, из которых 743 сигнала принадлежало больным, а 1055 сигналов принадлежало здоровым пациентам. Таким образом, нам необходимо ввести критерий качества признака.

Оценка качества синтезированного множества признаков будет происходить с помощью измерения качества алгоритма, обученного на этих признаках. Для этого может использоваться любой классический классификатор. В качестве базового классификатора был выбран случайный лес [2]. Это ансамбль решающих деревьев. Каждое решающее дерево строится по случайным подвыборкам, полученным в результате сэмплирований с возвращениями объектов обучающей выборки.

Для оценки качества будем использовать 20-кратную кросс-валидацию. Важно отметить, что в обучающей выборке многим пациентам принадлежит сразу несколько кардиограмм, поэтому валидация проводилась таким образом, чтобы кардиограммы любого пациента не могли попасть и в обучение, и в контроль одновременно. Это более честная оценка, так как кардиограммы одного и того же пациента очень похожи, и алгоритму проще выдать правильный ответ, так как он уже ранее видел похожую кардиограмму.

Размер случайной подвыборки составлял 100 объектов (50 объектов каждого класса). Функционал качества классификации – точность предсказания по пациентам. Вычисляется он так: для каждой кардиограммы каждого пациента делается предсказания о наличие у пациента болезни, затем для каждого пациента считается процент правильно классифицированных его кардиограмм, затем все эти значения усредняются по пациентам.

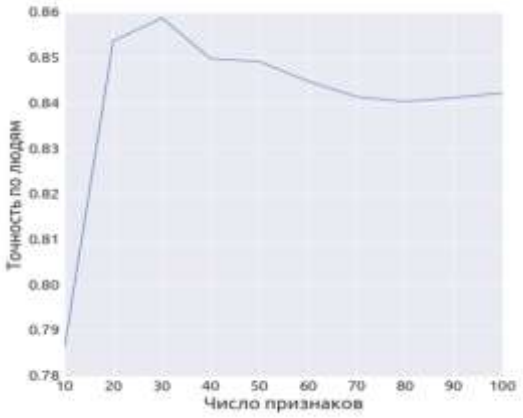


Рисунок 1 Зависимость точности по пациентам от количества признаков

Все эксперименты проводились с алгоритмом «случайный лес», который состоял из 100 деревьев. На Рис. 1 показана зависимость точности от количества признаков с шагом в 10 признаков. Видно, что оптимальное количество признаков находится около 30. Дальнейшее увеличение признакового пространства не приводит к росту качества. Это свидетельствует о том, что многие из сгенерированных признаков являются шумовыми. Важно отметить, что алгоритм постоянно создает новые признаки, они не совпадают с уже построенными. Максимальное значение точности по пациентам – 0.859.

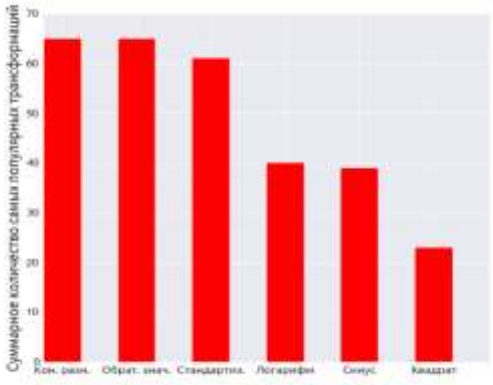


Рисунок 2 Самые популярные функции трансформации

Все дальнейшие эксперименты проводились для множества, состоящего из 300 синтезированных признаков. На рисунках 2 и 3 показаны самые часто встречаемые функции трансформации и агрегации. Как видно из этих рисунков, среди функций трансформаций нет определенной доминирующей функции, среди функций агрегаций с большим отрывом выигрывает медиана.

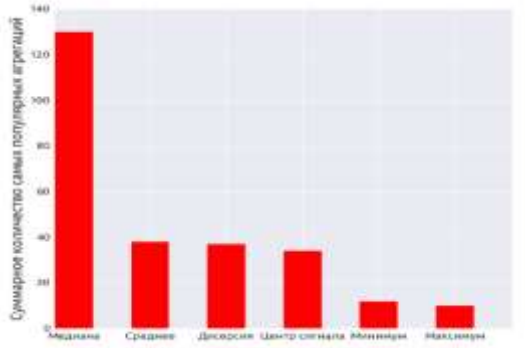


Рисунок 3 Самые популярные функции агрегации

Рисунки 4 и 5 показывают различие в поведении жадного алгоритма при различных методах оценки качества признака. Обозначения: ДНРП – доля неверно ранжированных пар (целевая переменная сортируется по значению признака), качество по ДР – качество признака по оценке дерева решений (см. раздел 2.2 про эти и другие оценки качества). Процент увеличения качества считается по формуле $\frac{FinalQual-InitQual}{InitQual}$, где $FinalQual$ – финальное качество признака, $InitQual$ – начальное качество. Начальное качество определяется качеством лучшей функцией агрегации при отсутствии функций трансформации.

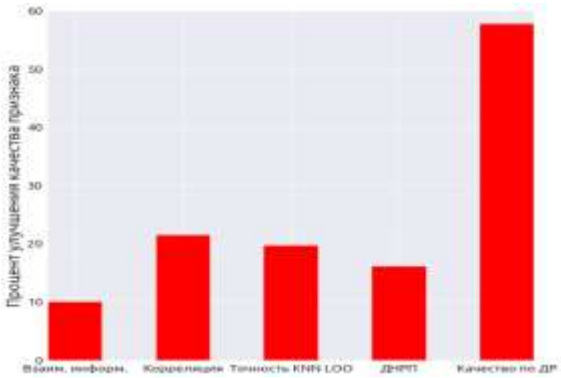


Рисунок 4 Средний процент увеличения качества признака

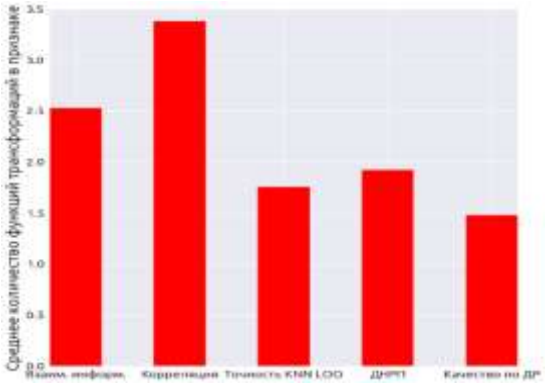


Рисунок 5 Средняя длина трансформаций в признаке

4 Выводы

Предложен алгоритм автоматического построения признакового пространства, проведены вычислительные эксперименты для задачи бинарной классификации кардиограмм. Данный алгоритм в вычислительных экспериментах показал свою способность конструировать признаковое пространство, которое позволило бы решать задачу классификации сигналов с высокой точностью. Перечислим основные достоинства данного подхода.

Алгоритм создает нужные признаки, используя только оценки качества этих признаков. Работа исследователя заключается только в выборе базисных функций, которые специфичны в его задаче. Например, исследователь может использовать фильтр, хорошо работающий конкретно для одного типа данных, но для сигналов этот фильтр не применим.

Алгоритм состоит из нескольких отдельных частей: начальный набор базисных функций, метод оценки качества признака, оптимизатор. Те варианты модулей, которые были приведены в данной работе, являются не более чем тестовыми вариантами, для каждой задачи они могут подбираться индивидуально.

Возможности алгоритма не ограничиваются его применением исключительно в задаче классификации сигналов. При изменении функций инициализации, трансформации и агрегации он может быть применен в любой дугой задаче распознавания неструктурированных данных, например, в задаче классификации текстов или изображений.

Благодаря своей стохастической природе алгоритм с каждой новой итерацией создает признак, который сильно отличается по своему методу построения от предыдущих. Чем больше итераций проведет алгоритм, тем больше вероятность, что среди полученных признаков будет подмножество действительно качественных.

Более подробное описание проблемы выделения признаков в задаче классификации сигналов можно найти в [9]. Настоящая работа содержит наиболее важные результаты вышеупомянутой.

Литература

[1] Al-Sahaf, H., Neshatian, K., Zhang, M.: Automatic Feature Extraction and Image Classification using Genetic Programming. In The 5th Int. Conf. on Automation, Robotics and Applications, pp. 157-162 (2011)

[2] Breiman, L.: Random Forests. Machine Learning (2001)

[3] Dal Seno, B., Matteucci, M., Mainardi, L.: A Genetic Algorithm for Automatic Feature Extraction in p300 Detection. 2008 IEEE Int. Joint Conf. on Neural Networks (IEEE World Congress

on Computational Intelligence), pp. 3145-3152 (2008)

[4] Guo, L., Rivero, D., Dorado, J., Munteanu, C., Pazos, A.: Automatic Feature Extraction using Genetic Programming: An Application to Epileptic Eeg Classification. Experts Systems with Applications: An Int. J. (2011)

[5] Kohler, B., Hennig, C., Orglmeister, R.: The Principles of Software qrs Detection. IEEE Engineering in Medicine and Biology Magazine (2002)

[6] Lallo, P.R.U.: Signal Classification by Discrete Fourier Transform. MILCOM 1999. IEEE Military Communications, pp. 197-201 (1999)

[7] Morik, K., Mierswa, K.: Automatic Feature Extraction for Classifying Audio Data. Machine Learning (2005)

[8] Prochazka, A., Kukal, J., Vysata, O.: Wavelet Transform use for Feature Extraction and Eeg Signal Segments Classification. 2008 3rd Int. Symposium on Communications, Control and Signal Processing, pp. 719-722 (2008)

[9] Викулин, В.А.: Автоматическое выделение признаков в задаче классификации сигналов. ВМК, МГУ имени М.В. Ломоносова (2017). http://www.machinelearning.ru/wiki/images/3/37/CourseVikulin.pdf

Диссертационный семинар

PhD Workshop

Интеграция данных, разработка схемы базы данных

Data integration, database schema development

Подход к реализации методов разрешения сущностей в среде распределенных вычислений Hadoop/MapReduce

© М.Д. Ислентьев

Московский государственный университет им. М.В. Ломоносова,
Москва, Россия

mikhail.islentyev@gmail.com

Аннотация. Процесс разрешения сущностей является частью процесса интеграции данных. Разрешение сущностей зависит, кроме всего прочего, от выбора мер сходства значений и методов сравнений пар записей. Данная статья посвящена разработке подхода к реализации мер сходства строковых значений как наиболее распространенных типов данных, а также детерминированных методов сравнения пар записей в среде распределенных вычислений Hadoop/MapReduce с использованием языка высокого уровня Jaql.

Ключевые слова: большие данные, интеграция данных, разрешение сущностей, связывание записей, удаление дубликатов, Hadoop, MapReduce, Jaql.

An Approach for Implementation of Methods for Entity Resolution in the Hadoop/MapReduce Distributed Computing Environment

© Mikhail D. Islentyev

Lomonosov Moscow State University,
Moscow, Russia

mikhail.islentyev@gmail.com

Abstract. The process of entity resolution is part of the data integration process. Entity resolution depends on the choice of similarity measures for values and methods for comparing pairs of records. In this paper, we develop an approach to implementation similarity measures of string values as the most common types of data, as well as deterministic methods for comparing pairs of records in the Hadoop/MapReduce distributed computing environment using the high-level language Jaql.

Keywords: big data, data integration, entity resolution, record linkage, data deduplication, Hadoop, MapReduce, Jaql.

1 Введение

В различных областях науки наблюдается экспоненциальный рост объема получаемых экспериментальных данных. Сложность использования таких данных увеличивается еще и вследствие их естественной разнородности. Это неизбежно приводит к необходимости использования неоднородной, распределенной информации, накопленной в течение значительного периода наблюдений различными инструментами.

Для анализа больших объемов данных используются современные среды распределенных вычислений, такие, как Hadoop/MapReduce [14, 19]. Такие среды имеют почти линейную горизонтальную масштабируемость и высокую отказоустойчивость. Основным достоинством подобных сред является возможность анализировать и обрабатывать разно-структурированные данные (реляционные, JSON, XML, тексты и др.). При этом возникает проблема интеграции данных, извлекаемых из разно-структурированных источников. Традиционно процесс интеграции данных включает в себя следующие шаги: унификация моделей данных, сопоставление схем, разрешение сущностей [6, 16] и слияние данных [5]. Остановимся подробней на этапе разрешения сущностей.

Данный этап нацелен на поиск записей в одном или нескольких наборах данных, представляющих собой один и тот же объект в реальном мире, или сущность. Он ориентирован на решение таких задач, как связывание записей, выявление и удаление дубликатов, сопоставление связей и др.

Труды XIX Международной конференции «Аналитика и управление данными в областях с интенсивным использованием данных» (DAMDID/ RCDL'2017), Москва, Россия, 10–13 октября 2017 года

Весь процесс разрешения сущностей, в соответствии с [10], можно разделить на следующие этапы: подготовка данных; выбор мер сходства значений; выбор метода сравнения пар записей; определение ограничений.

Данная работа сосредоточена на втором и третьем этапах. Выбор мер сходства является одним из самых важных этапов в процессе разрешения сущностей. Важно выбрать меры, наиболее подходящие для представленного набора данных, поскольку именно на основе значений этих мер будет делаться вывод, являются ли записи в паре совпадающими, то есть относящимися к одной сущности, или же нет. И поскольку большинство данных представлено в виде строковых значений (имена, названия, адреса и т. д.), особое внимание уделяется нами мерам сходства строк. Среди методов сравнения пар записей известны детерминированные методы, к которым относятся метод взвешенной суммы и метод, основанный на формулировании правил. Кроме того, отдельно стоят методы разделения на блоки. Они не входят в традиционный процесс разрешения сущностей, однако могут значительно увеличить скорость выполнения и производительность всего процесса при помощи эффективного создания небольших блоков, содержащих только потенциально совпадающие записи.

Нашей целью является разработка подхода к реализации методов разрешения сущностей в среде распределенных вычислений Hadoop/MapReduce. Для непосредственной реализации мер и методов в среде Hadoop/MapReduce был выбран язык высокого уровня Jaql [4]. Jaql – это функциональный декларативный язык запросов, который предназначен для обработки больших наборов данных. Он позволяет работать с разно-структурированными данными, распределенной файловой системой HDFS и, при необходимости, сам переписывает высокоуровневые запросы в запросы «низкого уровня», состоящие из MapReduce-задач.

На языке Jaql реализован ряд мер сходства строковых значений, алгоритмов сравнения пар записей и разбиения записей на блоки, рассмотрены некоторые идеи и особенности реализации. Выбранный набор мер сходства строк обширен, и в большинстве случаев его достаточно для задач разрешения сущностей [17], однако реализация позволяет определять собственные меры сходства, в том числе и для значений, не являющимися строковыми, в виде непосредственно функций на Jaql или же в виде Java UDF (пользовательских функций Java). Код реализации доступен в git-репозитории [12].

В разделе 2 описаны меры сходства строковых значений, приводятся примеры использования и реализации некоторых мер и примеры задания и использования компаратора для получения вектора сравнения пары записей. В разделе 3 описаны детерминированные методы сравнения пар записей и показаны примеры их использования. В разделе 4 обсуждается тема разбиения на блоки для уменьшения числа попарных сравнений и приведены примеры их применения. Наконец, в разделе 5 показан пример полного процесса разрешения сущностей в рамках реализованных мер и методов.

2 Меры сходства значений

Ключевым моментом в процессе разрешения сущностей является получение оценок сходства значений соответствующих атрибутов двух сравниваемых записей путем вычисления мер их сходства. Полученные оценки формируют вектор сравнения, на основе которого в дальнейшем делаются выводы о совпадении или различии рассматриваемой пары записей. Таким образом, важно выбрать наиболее подходящие под имеющиеся данные меры сходства.

Подавляющее большинство значений атрибутов представляет собой строки, и, кроме того, верное определение сходства строк не всегда является тривиальным действием, поэтому особое внимание уделим мерам сходства именно строковых значений. По своей природе их обычно делятся на следующие группы:

- меры сходства на основе редактирования;
- меры сходства на основе разбиения на токены;
- гибридные меры сходства.

Рассмотрим подробнее каждую группу.

2.1 Меры сходства на основе редактирования

Меры данного типа оперируют числом вставок, удалений, замен и/или перестановок символов в строке. Чем больше требуется таких операций для преобразования одной строки к другой, тем менее они похожи.

Для реализации мер этой группы в основном применяется метод динамического программирования [11, 18]. Декларативные языки программирования, к которым относится Jaql, не предназначены для реализации подобных методов, поэтому было решено воспользоваться расширяемостью языка при помощи Java UDF.

Каждая такая функция принимает два строковых параметра (см. Программу 1). В данном примере показан вызов функции расстояния Джаро–Винклера [7] для строк “Dwayne” и “Duane”.

Программа 1 Вызов функции меры сходства на основе редактирования на примере расстояния Джаро–Винклера

```
import similarity;
similarity::jaroWinklerSim("Dwayne",
"Duane");
```

Реализованные меры на основе редактирования: расстояние Левенштейна [7], расстояние Дамерау–Левенштейна [7], наибольшая общая подпоследовательность [11], расстояние Джаро [7], расстояние Джаро–Винклера.

2.2 Меры сходства на основе разбиения на токены

Принадлежащие к этой группе меры используют разбиение строк на токены. Чаще всего применяют

два вида разбиения: разбиение на слова и разбиение на n-граммы.

Для расчета меры сходства набор токенов представляют либо в качестве множества, и тогда применяют меры сходства множеств, либо в качестве вектора в многомерном пространстве, после чего используют меры сходства векторов.

Реализация данных мер подразумевает нахождение пересечения двух наборов токенов, что просто и эффективно реализуется на Jaql с помощью встроенной функции join (см. Программу 2). В качестве примера приведем реализацию коэффициента Дайса [7]:

$$sim_{Dice}(x, y) = \frac{2n_t}{n_x + n_y},$$

где x, y – сравниваемые строки, n_x, n_y – число токенов в строке x и y соответственно, n_t – число совпадающих токенов в строках x и y.

Программа 2 Реализации функции меры сходства на основе разбиения на токены на примере коэффициента Дайса

```
bagsIntersection = fn(lhs, rhs)
  join lhs, rhs
  where lhs.token == rhs.token
  into {
    lhs.token,
    count: min([ lhs.count, rhs.count ])
  };

diceSim = fn(lhs, rhs)
  count(bagsIntersection(lhs, rhs)) *
  2.0 / (count(lhs) + count(rhs));
```

Функции мер данной группы принимают строки, предварительно разбитые на токены.

Список реализованных мер сходства на основе разбиения на токены: коэффициент Дайса, коэффициент Жаккара [7], коэффициент перекрытия [7], косинусный коэффициент [7], статистическая мера TFIDF [7].

2.3 Гибридные меры сходства

К таким мерам сходства относятся меры, оперирующие наборами токенов и применяющие к сравнению токенов меры на основе редактирования. Такие меры отличают повышенная точность оценивания сходства, но в то же время увеличенное время выполнения [7]. Поскольку названные меры также используют наборы токенов, реализация этих методов тоже была выполнена на языке Jaql. Были реализованы следующие гибридные меры сходства: сходство Монг-Элкана [7] и статистическая мера SoftTFIDF [7].

2.4 Вектор сравнения

Вектор сравнения, как было сказано ранее, формируется из оценок сходства для значений атрибутов пар записей. В реализации используется компаратор в виде записи, атрибутами которой являются имена оценок (используются далее в методах сравнения пар записей), а значениями – функции мер сходства (см. Программу 3).

Программа 3 Пример задания компаратора

```
import similarity;
addressComparator = fn(lhs, rhs)
  similarity::jaccardSim(
    lhs.address -> similarity::nGramBag(),
    rhs.address -> similarity::nGramBag()
  );
nameComparator = fn(lhs, rhs)
  similarity::minLengthMongeElkanSim(
    lhs.name -> similarity::wordBag(),
    rhs.name -> similarity::wordBag()
  );
typeComparator = fn(lhs, rhs)
  similarity::equalSim(lhs.type, rhs.type);

recordComparator = {
  addressScore: addressComparator,
  nameScore: nameComparator,
  typeScore: typeComparator
};
```

Далее этот компаратор может быть передан функции вместе с парой записей, для которых необходимо вычислить вектор сравнения (см. Программу 4).

Программа 4 Пример получения вектора сравнения

```
import resolution;
pair = [ { ... }, { ... } ];
vector = recordComparator
  -> resolution::countVector(pair);
```

Реализация функции countVector() получает от компаратора пары атрибут/значение и создает вектор сравнения – запись с теми же именами атрибутов, значения которых являются оценками сходства (см. Программу 5).

Программа 5 Реализация функции countVector()

```
countVector = fn(recordComparator, pair)
  recordComparator -> fields()
    -> transform {
      ($[0]): _evalFieldComparator(
        $[1], pair[0], pair[1]
      )
    } -> record();
```

Закрытая функция _evalFieldComparator() выполнена в виде Java UDF и просто вызывает переданную функцию меры для пары записей. Такой подход к реализации позволил сохранить возможность выполнения данной функции внутри MapReduce-задачи, поскольку явный вызов функции через индексатор ([1]) не позволяет движку оптимизации Jaql создать MapReduce-задачу.

3 Детерминированные методы сравнения пар записей

Пусть имеются пара записей и их вектор сравнения, полученный при помощи компаратора, объявленного ранее:

```
vector = { addressScore, nameScore, typeScore
}
```

Возникает проблема, как по вектору сравнений определить, являются ли записи совпадающими или нет.

3.1 Взвешенная сумма

Наиболее простым решением этой проблемы

является использование среднего значения всех оценок сходства из вектора сравнений или же их взвешенной суммы, после чего необходимо задать пороговое значение, определяющее совпадение или различие записей, например (веса и порог здесь и далее выбраны случайно):

```
0.2 * addressScore + 0.8 * nameScore > 0.85
```

В реализации данного метода вектор весов также представляет собой запись с теми же именами атрибутов, что и у компаратора, значения которого и являются весами (см. Программу 6).

Программа 6 Пример метода взвешенной суммы

```
import
  resolution::classifier::weightedSum as ws;
vector -> ws::classifier({
    addressScore: 0.2,
    nameScore: 0.8
  },
  0.85
);
```

3.2 Правила

Несколько иным способом является формулирование набора правил, где ограничения накладываются на оценки сходства каждого атрибута независимо. Пример таких правил (оператор ‘&’ обозначает логическое И, оператор ‘|’ — логическое ИЛИ):

```
typeScore = 1.0 &
(nameScore > 0.7 | addressScore > 0.9) |
nameScore > 0.9
```

Метод сравнения, использующий данные правила, считает записи в паре совпадающими, если поле `nameScore` их вектора сравнений больше 0.9, или же если `typeScore` равно 1.0, и при этом либо `nameScore` больше 0.7, либо `addressScore` больше 0.9.

Правила реализуются записью, представляющей собой двоичное дерево выражений. В следующем примере составлено дерево, эквивалентное описанным выше правилам (см. Программу 7).

Программа 7 Пример метода на основе правил

```
import resolution::classifier::ruleBased as
rb;
rules = {
  "lhs": {
    "lhs": {
      "lhs": "typeScore",
      "op": "=",
      "rhs": 1.0
    },
    "op": "&",
    "rhs": {
      "lhs": {
        "lhs": "nameScore",
        "op": ">",
        "rhs": 0.7
      },
      "op": "|",
      "rhs": {
        "lhs": "addressScore",
        "op": ">",
        "rhs": 0.9
      }}},
  "op": "|",
  "rhs": {
    "lhs": "nameScore",
    "op": ">",
    "rhs": 0.9
  }};
vector -> rb::classifier(rules);
```

Даже такое небольшое количество правил требует построения достаточно громоздкого двоичного дерева выражений. Для облегчения формулирования и применения правил был реализован синтаксический анализатор в виде Java UDF, позволяющий получить дерево выражений из строкового выражения (см. Программу 8).

Программа 8 Пример разбора строкового выражения правил

```
import resolution::classifier::ruleBased as
rb;
rb::parseRules(
  "typeScore = 1.0 & " +
  "(nameScore > 0.7 | addressScore > 0.9) |
" +
  "nameScore > 0.9"
);
```

Результатом данного вызова функции будет дерево, эквивалентное вышеописанному.

4 Методы разделения на блоки

С увеличением числа данных полное попарное сравнение становится крайне неэффективным. Действительно, полное попарное сравнение набора данных, состоящих из n записей, потребует $n(n-1)/2$, или $O(n^2)$, сравнений. Для уменьшения числа пар записей, которые необходимо сравнить, используются методы разделения на блоки. Такие методы отвергают заведомо несовпадающие пары записей и создают блоки, состоящие из пар, которые потенциально могут совпадать.

Для рассмотрения были выделены следующие методы: метод исключительного разделения на блоки [13], метод индексации биграмм [2] и метод на основе кластеризации с помощью canopy [15]. Реализация этих методов разделена на две части: одна – для задачи выявления и удаления дубликатов (в одном наборе данных), вторая – для задачи связывания записей (для двух наборов данных).

4.1 Метод исключительного разделения

Самым простым является метод исключительного разделения на блоки. Идея метода состоит в том, что набор данных делится на непересекающиеся блоки по блочному ключу. Такой ключ может являться значением какого-либо атрибута записи или же комбинацией нескольких значений или даже их частью.

Реализация данного метода подразумевает группировку записей по блочному ключу, что в языке Jaql возможно совершить с помощью встроенной функции group by. Применение такого метода требует задания функции, генерирующей блочный ключ (см. Программу 9).

Программа 9 Пример метода исключительного разделения для двух наборов данных

```
import
resolution::linkage::blocking::simple;
```

```
data1 = read(...);
data2 = read(...);
data1Key = fn(value)
  value.lastname -> substring(0, 4);
data2Key = fn(value)
  value.surname -> substring(0, 4);

simple::blocking(data1, data2,
  data1Key, data2Key
);
```

Преимуществом этого метода является высокая скорость его исполнения. К недостаткам же можно отнести сложность выбора критерия, а также то, что при не совсем оптимальном его выборе реально совпадающие записи могут попасть в различные блоки, тем самым они никогда не будут сравниваться.

4.2 Индексация биграмм

Следующий метод, называемый методом индексации биграмм, позволяет осуществлять приближенное разделение на блоки. Его алгоритм описан в [2]. Реализация метода сопряжена с получением инвертированного индекса и группировкой, с чем Jaql прекрасно справляется, используя встроенные функции group by и expand unroll. Для проведения такого разбиения необходимо помимо функции блочного ключа подать пороговое значение (см. Программу 11).

Программа 11 Пример метода индексации биграмм для одного набора данных

```
import

resolution::deduplication::blocking::bigram;
data = read(...);
dataKey = fn(value)
  value.address
    -> strSplit("\\s+") -> index(0);
data -> bigram::blocking(dataKey, 0.3);
```

4.3 Кластеризация canopy

Иной подход к разделению на блоки реализует метод, основанный на кластеризации с помощью canopy. Данный метод кластеризации опирается на возможность для случайной записи из набора данных эффективно найти все близлежащие записи при помощи какой-либо приближенной функции расстояния, требующей небольших вычислительных затрат. После проведения кластеризации каждый canopy-кластер формирует свой блок.

Данный алгоритм кластеризации имеет вариант реализации непосредственно на MapReduce, поэтому реализован он был с помощью явного задания MapReduce-задачи на языке Jaql. Для применения метода разделения, основанного на такой кластеризации, необходимо задать функцию расстояния и два пороговых значения (см. Программу 12).

Программа 12 Пример метода кластеризации canopy для одного набора данных

```
import similarity;
import

resolution::deduplication::blocking::canopy;
data = read(...);
distanceFunction = fn(lhs, rhs)
  1.0 - similarity::minLengthMongeElkanSim(
    lhs.name -> similarity::wordBag(),
    rhs.name -> similarity::wordBag()
  );
data -> canopy::blocking(
  distanceFunction, 0.08, 0.16
);
```

Метод индексации биграмм и метод на основе кластеризации с помощью canopy генерируют пересекающиеся блоки, что снижает вероятность разделения на разные блоки совпадающих записей. Также эти методы имеют хороший и близкий друг к другу результат по качеству разделения при правильном выборе функций и порогов, что отражено в [2].

4 Пример использования

Для примера рассмотрим применение процесса разрешения сущностей для одного набора данных (см. Программу 13).

Программа 13 Пример процесса разрешения сущностей для одного набора данных

```
import similarity;
import
  resolution::classifiers::weightedSum    as
ws;
import resolution::deduplication;
import

resolution::deduplication::blocking::canopy;

data = read(...);
distanceFunction = fn(lhs, rhs)
  1.0 - similarity::minLengthMongeElkanSim(
    lhs.name -> similarity::wordBag(),
    rhs.name -> similarity::wordBag()
  );
blockingInfo = canopy::createInfo(
  distanceFunction, 0.08, 0.16
);
addressComparator = fn(lhs, rhs)
  similarity::jaccardSim(
    lhs.address -> similarity::nGramBag(),
    rhs.address -> similarity::nGramBag()
  );
nameComparator = fn(lhs, rhs)
  similarity::minLengthMongeElkanSim(
    lhs.name -> similarity::wordBag(),
    rhs.name -> similarity::wordBag()
  );
typeComparator = fn(lhs, rhs)
  similarity::equalSim(lhs.type, rhs.type);

recordComparator = {
  addressScore: addressComparator,
  nameScore: nameComparator,
  typeScore: typeComparator
};

classifierInfo = ws::createInfo({
    addressScore: 0.2,
    nameScore: 0.8
  },
  0.85
);
data                                        ->
deduplication::deduplicateWithBlocking(
  blockingInfo,
  recordComparator,
  classifierInfo
);
```

Данные после чтения будут разделены на блоки с помощью метода кластеризации canopy, затем блоки преобразуются в пары записей, для каждой такой пары будет вычислен их вектор сравнения на основе трех переданных функций мер, по этому вектору метод взвешенной суммы определит, являются ли записи в паре совпадающими или нет, после чего будут возвращены только те пары, записи в которых были определены как совпадающие.

5 Заключение и дальнейшая работа

Мы описали меры сходства строковых значений, детерминированные методы сравнения пар записей и разработали подход к их реализации в среде распределенных вычислений Hadoop/MapReduce с использованием высокоуровневого языка программирования Jaql. Также описаны и реализованы методы по разделению пар записей на блоки для значительного снижения числа попарных сравнений и увеличения производительности.

Набор реализованных функций мер сходства строковых значений достаточно обширен, однако реализация позволяет определять собственные меры сходства, в том числе и для значений, не являющимися строковыми, в виде непосредственно функций на Jaql или же в виде Java UDF.

В дальнейшем планируется реализовать поддержку ограничений [10] и метод корреляционной кластеризации [8] для задачи выявления и удаления дубликатов, а также рассмотреть возможность реализации иных методов сравнения пар записей в среде Hadoop, таких, как вероятностные методы [9] и методы, основанные на машинном обучении [1, 3].

Поддержка

Работа выполнена при поддержке РФФИ (гранты 15-29-06045, 16-07-01028).

Литература

[1] Arasu, A., Götz, M., Kaushik, R.: On Active Learning of Record Matching Packages. Proc. of the 2010 ACM SIGMOD Int. Conf. on Management of Data, pp. 783-794. ACM, New York (2010)

[2] Baxter, R., Christen, P., Churches, T.: A Comparison of Fast Blocking Methods for Record Linkage. Proc. of the KDD-03 Workshop on Data Cleaning, Record Linkage and Object Consolidation, pp. 25-27. ACM, New York (2003)

[3] Bellare, K., Iyengar, S., Parameswaran, A.G., Rastogi, V.: Active Sampling for Entity Matching. Proc. of the 18th ACM SIGKDD Int. Conf. on Knowledge Discovery and Data Mining, pp. 1131-1139. ACM, New York (2012)

[4] Beyer, K.S., Ercegovac, V., Gemulla, R., Balmin A., Eltabakh, M.Y., Kanne, C.C., Özcan, F., Shekita, E.J.: Jaql: A Scripting Language for Large Scale Semistructured Data Analysis. Proc. of the VLDB Endowment, 4 (12), pp. 1272-1283 (2011)

[5] Bleiholder, J., Naumann, F.: Data Fusion. ACM Computing Surveys, 41 (1), pp. 1:1–1:41 (2009)

[6] Christen, P.: Data Matching – Concepts and Techniques for Record Linkage, Entity Resolution, and Duplicate Detection. Springer, Heidelberg (2012)

[7] Cohen, W.W., Ravikumar, P., Fienberg, S.E.: A Comparison of String Distance Metrics for Name-Matching Tasks. Proc. of the 2003 Int. Conf. on Information Integration on the Web, pp. 73-78. AAAI Press (2003)

[8] Elsner, M., Schudy, W.: Bounding and Comparing Methods for Correlation Clustering Beyond ILP. Proc. of the Workshop on Integer Linear Programming for Natural Language Processing, pp. 19-27. Association for Computational Linguistics, Stroudsburg (2009)

[9] Fellegi, I.; Sunter, A.: A Theory for Record Linkage. J. of the American Statistical Association, 64 (328), pp. 1183–1210 (1969)

[10] Getoor, L., Machanavajjhala, A.: Entity Resolution for Big Data. Proc. of the 19th ACM SIGKDD Conf. on Knowledge Discovery and Data Mining, p. 1527. ACM, New York (2013)

[11] Hirschberg, D.S.: A Linear Space Algorithm for Computing Maximal Common Subsequences. Communications of the ACM, 18 (6), pp. 341-343 (1975)

[12] Islentyev, M.D.: mislen/jaql-entity-resolution: Entity Resolution Methods on Jaql (2017). https://github.com/mislen/jaql-entity-resolution

[13] Jaro, M.A.: Advances in Record Linkage Methodology as Applied to Matching the 1985 Census of Tampa, Florida. J. of the American Statistical Association, 84 (406), pp. 414-420 (1989)

[14] Maitrey, S., Jha, C.K.: MapReduce: Simplified Data Analysis of Big Data. Procedia Computer Science, 57, pp. 563-571 (2015)

[15] McCallum, A., Nigam, K., Ungar, L.H.: Efficient Clustering of High-Dimensional Data Sets with Application to Reference Matching. Proc. of the 6th ACM SIGKDD Int. Conf. on Knowledge Discovery and Data Mining, pp. 169-178. ACM, New York (2000)

[16] Naumann, F., Herschel, M.: An Introduction to Duplicate Detection. Morgan and Claypool Publishers (2010)

[17] Vovchenko, A.E., Kalinichenko, L.A., Kovalev, D.Y.: Methods of Entity Resolution and Data Fusion in the ETL-Process and their Implementation in the Hadoop Environment. Informatics and Applications, 8 (4), pp. 94-109 (2014)

[18] Wagner, R.A.; Fischer, M.J. The String-to-String Correction Problem. J. of the ACM, 21 (1), pp. 168-173 (1974)

[19] White, T.: Hadoop: The Definitive Guide, 4th Edition. O'Reilly Media (2015)

Построение схем реляционных баз данных с помощью элементарных связей атрибутов: алгоритм вычисления замыкания атрибутов для одного типа связи

© И.П. Убалехт

Омский государственный технический университет,
Омск, Россия

ivan@ubaleht.com

Аннотация. В рамках разрабатываемого метода построения схем реляционных баз данных введены понятия «элементарная связь атрибутов» и «тип элементарной связи атрибутов». Для одного из введённых типов связи – связи типа 1..M:0..1, построена система правил вывода, предложен алгоритм построения замыкания атрибутов и membership-алгоритм. Доказана корректность работы алгоритма построения замыкания атрибутов относительно множества связей типа 1..M:0..1. Предложенный membership алгоритм позволяет определить принадлежность произвольной связи типа 1..M:0..1 замыканию исходного множества связей типа 1..M:0..1.

Ключевые слова: Построение схем баз данных, модели данных, ограничения целостности.

Design of Relational Database Schemes Based on the Elementary Relationships of Attributes: Algorithm of Computation Closure of a Set of Attributes for One Type of Relationship

© Ivan Ubaleht

Omsk State Technical University,
Omsk, Russia

ivan@ubaleht.com

Abstract. We present following concepts: “elementary relationships of attributes”, “type of elementary relationship of attributes”. These concepts are used for the method of design of relational database schemes. We present special case, in which initial set of relationships consists of only relationships of 1..M:0..1 type. For this special case we propose: the set of inference rules, algorithm of computing of the closure of a set attributes with respect to a set of relationships of 1..M:0..1 type and algorithm to test membership in the closure of elementary relationships of attributes for some relationship of 1..M:0..1 type. Furthermore, we propose a proof of the correctness of the algorithm of computing of the closure of a set attributes with respect to a set of relationships of 1..M:0..1 type.

Keywords: design of schemes of relational databases, schemas of relational databases, data models.

1 Введение

В настоящее время остаётся актуальной задача разработки методов построения схем реляционных баз данных (РБД), обладающих высокой степенью формализации и автоматизации процесса формирования логических схем РБД, а также обеспечивающих эффективное взаимодействие с пользователем-проектировщиком схем РБД. Наиболее известны следующие подходы к получению оптимального логического дизайна РБД:

- использование модели Сущность – Связь (ER-модель), задействован инфологический уровень;

- разработка логических схем РБД с применением теории нормальных форм, с применением декомпозиции отношений (метод декомпозиции) [12, 15], реже – синтеза (метод синтеза); под методом синтеза будем понимать группу методов, в соответствии с которыми схемы БД формируются из начального множества функциональных зависимостей (ФЗ) и атрибутов [12, 15].

Существуют и другие менее распространённые

Труды XIX Международной конференции «Аналитика и управление данными в областях с интенсивным использованием данных» (DAMDID/ RCDL’2017), Москва, Россия, 10–13 октября 2017 года

подходы, используемые при разработке проекта БД, в том числе, на стадии логического проектирования. Например, использование модифицированных вариантов ER-модели [4, 14], использование ORM-модели вместо ER-модели [6, 7], построение схем БД с помощью языков логического программирования [9], системы, позволяющие осуществлять настройку (tuning) РБД автономно на логическом уровне, учитывая текущую нагрузку на РБД [3], а также множество других подходов, проектов, программных утилит [1, 5, 11, 13, 16].

В данной статье в рамках разрабатываемого метод построения схем РБД предложена концепция элементарных связей атрибутов (далее ЭСА или просто связи). ЭСА являются элементарными утверждениями о предметной области (ПрО). Практически любое высказывание о количественном взаимоотношении двух характеристик некоторого объекта в ПрО можно свести к ЭСА определённого типа. Например, утверждение из некоторой ПрО: «Для заданного табельного номера инженера (*Таб_Номер_Инж*) имеется строго одно наименование категории инженера (*Категория_Инж*) и для заданного одного наименование категории инженера имеется не менее одного табельного номера инженера». Данное утверждение из ПрО можно свести к следующей компактной форме записи в виде ЭСА: *Таб_Номер_Инж* $\xleftarrow{1..M:1..1}$ *Категория_Инж*. Как видно, концепция ЭСА является вариантом формализации понятия cardinality constraint. В данной статье для частного случая – для набора ЭСА типа 1..M:0..1 ставятся задачи: предложить правила вывода для ЭСА типа 1..M:0..1; на основе правил вывода разработать алгоритм вычисления замыкания атрибутов относительно исходного множества ЭСА типа 1..M:0..1; предложить membership-алгоритм, определяющий принадлежность произвольной связи типа 1..M:0..1 замыканию связей типа 1..M:0..1. Решение этих задач необходимо для получения минимальных покрытий и построения схем РБД методом синтеза и для разработки алгоритма построения схем РБД методом декомпозиции отношений.

2 Элементарные связи атрибутов

Дадим определение понятию «элементарная связь атрибутов».

Определение 1. Пусть $\boldsymbol{A}=\{A_1, .. A_n\}$ и $\boldsymbol{B}=\{B_1, .. B_m\}$ – множества атрибутов, где каждый атрибут из множеств $\boldsymbol{A}$ и $\boldsymbol{B}$ является именем домена, n – мощность $\boldsymbol{A}$, m – мощность $\boldsymbol{B}$, пусть $r_1(\boldsymbol{A})$ и $r_2(\boldsymbol{B})$ – отношения со схемами $\boldsymbol{A}$ и $\boldsymbol{B}$ соответственно. Тогда *связью RS*, заданной на $\boldsymbol{A}$ и $\boldsymbol{B}$, является

$$RS \subseteq (r_1(\boldsymbol{A}) \times r_2(\boldsymbol{B})) \cup r_3(\boldsymbol{AB}) \cup r_4(\boldsymbol{AB}),$$

где $r_3(\boldsymbol{AB})$ – множество кортежей вида $\{t(\boldsymbol{AB}) \mid a_1,..,a_n$ – значения принадлежащие доменам, обозначаемым атрибутами $A_1,..,A_n$; $b_1,..,b_m$ – выделенные значения, обозначаемые как *null*$\}$, $r_4(\boldsymbol{AB})$ – множество кортежей вида $\{t(\boldsymbol{AB})|a_1,..,a_n$ – выделенные значения, обозначаемые как *null*; $b_1,..,b_m$ – значения принадлежащие доменам, обозначаемым атрибутами $B_1,..,B_m\}$.

Определение 2. Пусть *RS* – связь, заданная на $\boldsymbol{AB}$, $r_1(\boldsymbol{A})$ и $r_2(\boldsymbol{B})$ – отношения на схемах $\boldsymbol{A}$ и $\boldsymbol{B}$, $r_3(\boldsymbol{AB})$ и $r_4(\boldsymbol{AB})$ – отношения, такие же, как в Определении 1. Правило получения множества кортежей

$$RS \subseteq (r_1(\boldsymbol{A}) \times r_2(\boldsymbol{B})) \cup r_3(\boldsymbol{AB}) \cup r_4(\boldsymbol{AB})$$

будем называть типом связи *RelShipType*. Связь *RS* типа *RelShipType* будем обозначать следующим образом – *RS*: $\boldsymbol{A} \xleftarrow{RelShipType} \boldsymbol{B}$. Множества атрибутов $\boldsymbol{A}$ и $\boldsymbol{B}$ будем называть сторонами связи *RS*.

Стороны связи примерно соответствуют понятиям детерминанта ФЗ и зависимой части ФЗ.

Для пояснения Определений 1 и 2 на Рис. 1 показан пример отношения, реализующего связь типа 0..M:0..1,

Таб_Номер_Инж $\xleftarrow{0..M:0..1}$ *Категория_Инж*,

т. е. все кортежи этого отношения не должны противоречить правилу, задающему тип 0..M:0..1.

Рисунок 1 Пример ЭСА типа 0..M:0..1, обозначения в соответствии с Определением 1

В Таблице 1 показано несколько типов связей. Правила, определяющее типы связей, заданы с помощью утверждений на естественном языке, ниже будет более строго формализован тип связи 1..M:0..1. Обозначения и определения типов связей из Таблицы 1 аналогичны обозначениям и определениям типов связей, предложенным в статье К. Дейта [2]. Как видно из Таблицы 1, данные типы связей могут быть поставлены в соответствие с часто встречаемыми на практике ограничениями целостности (см. третий столбец Таблицы 1).

Все элементарные связи атрибутов должны удовлетворять следующим структурным свойствам.

Структурное свойство 1. Пусть *RS*: $\boldsymbol{A} \xleftarrow{RelShipType} \boldsymbol{B}$ – связь, заданная на $\boldsymbol{AB}$. Для каждого кортежа $(t|d_1,..,d_k)$, принадлежащего $\pi_A(RS)$ (стороне связи $\boldsymbol{A}$) либо принадлежащего $\pi_B(RS)$ (стороне связи $\boldsymbol{B}$), выполнены условия:

1. все значения $d_1,..,d_k$ являются значениями, принадлежащими доменам из $\boldsymbol{A}$ или $\boldsymbol{B}$, и тогда такой кортеж будем называть *value-total* кортежем;

2. все значения $d_1,..,d_k$ являются null-значениями, и тогда такой кортеж будем называть *null-total* кортежем.

Следовательно, все кортежи, принадлежащие стороне какой-либо ЭСА, являются либо *value-total* либо *null-total*.

Таблица 1 Типы ЭСА и ограничения целостности

Типы ЭСА	Правила, определяющие тип связи, где $\boldsymbol{A}$ и $\boldsymbol{B}$ – множества атрибутов, элемент (кортеж) $a \in \boldsymbol{A}$, элемент (кортеж) $b \in \boldsymbol{B}$	Ограничения целостности
1..M:1..1	Для заданного a имеется строго один элемент b и для заданного b имеется не менее одного элемента a	Функциональные зависимости
1..1:0..1	Для заданного a имеется не более одного элемента b и для заданного b имеется строго один элемент a	
1..M:0..1	Для заданного a имеется не более одного элемента b и для заданного b имеется не менее одного элемента a	Отношение между первичным ключом и внешним ключом в одной и той же таблице (во внешнем ключе допускаются неопределённые значения)
0..1:1..1	Для заданного a имеется строго один элемент b и для заданного b имеется не более одного элемента a	
0..M:1..1	Для заданного a имеется строго один элемент b и для заданного b имеется M элементов a, где $M \geq 0$	Отношение между первичным ключом и внешним ключом для двух разных таблиц (зависимости включения)
0..1:0..1	Для заданного a имеется не более одного элемента b и для заданного b имеется не более одного элемента a	
0..M:0..1	Для заданного a имеется не более одного элемента b и для заданного b имеется M элементов a, где $M \geq 0$	Отношение между первичным ключом и внешним ключом для двух разных таблиц, когда во внешнем ключе допускаются неопределённые значения (этот случай не полностью отражается зависимостями включения)
1..M:1..1	Для заданного a имеется строго один элемент b и для заданного b имеется не менее одного элемента a	

Структурное свойство 2. Пусть RS: $\boldsymbol{A} \xleftarrow{RelShipType} \boldsymbol{B}$ – элементарная связь атрибутов. Для любых двух кортежей t_1 и $t_2 \in RS$, таких, что $\pi_A(t_1)=\pi_A(t_2)$, выполняется следующее: $\pi_B(t_1)$ и $\pi_B(t_2)$ будут либо *value-total*, либо *null-total*.

В приведенных выше определениях используются null-значения. В теории РБД существует множество различных интерпретаций неопределённых значений и неполной информации [8, 10, 17]. В данной работе null-значения фактически не являются неопределёнными значениями в том смысле, в котором их понимают в приведённых выше источниках. Как видно из структурного свойства 1, null-значения всегда «заполняют» всю проекцию кортежа на сторону связи. Таким образом, связи, содержащие *null-total* кортежи, приближённо соответствуют связям между таблицами с необязательным классом принадлежности в ER-модели. Похожая интерпретация неопределённых значений и понятия «связь», но в менее формализованном виде, приведена в [2].

На Рис. 2 показаны случаи, не соответствующие структурным свойствам 1 и 2. Пусть в отношении на Рис. 2 атрибуты A и B составляют первичный ключ, а X и Y – внешний ключ с каким-то другим отношением. Если исключить кортежи, выделенные на Рис. 2, то взаимоотношение AB и XY можно выразить связью $AB \xleftarrow{1..M:1..1} XY$. Как видно из Рис. 2, в практике проектирования РБД, как правило, не бывает случаев, чтобы кортежи, составляющие внешний ключ, имели бы вид как у кортежей, выделенных в рамки на Рис. 2. Если бы null-значения интерпретировались одним из традиционных способов, то такой вид кортежей был бы теоретически возможен.

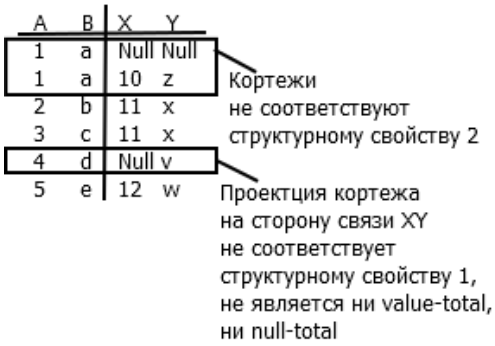


Рисунок 2 Примеры кортежей, не соответствующих структурным свойствам 1 и 2

В Таблице 1 представлено, каким ограничениям, важным для проектирования и работы РБД, соответствуют заданные типы ЭСА. Как видно, чтобы это соответствие соблюдалось, ЭСА должны отвечать структурным свойствам 1 и 2.

3 Частный случай. Связи типа 1..M:0..1

3.1 Элементарная связь атрибутов типа 1..M:0..1 и правила вывода

Как указано в Таблице 1, связь типа 1..M:1..1 соответствует ФЗ. Если проект РБД состоит только из связей типа 1..M:1..1 или 1..1:1..1 (которые можно свести к 1..M:1..1), то методы построения схем РБД сведутся к классическим подходам. Один из классических подходов – это использование метода синтеза. Важными элементами метода синтеза

являются: правила (аксиомы) вывода; замыкание множества ФЗ, основанное на аксиомах вывода; замыкание атрибутов; membership-алгоритм, определяющий принадлежность произвольной ФЗ заданному множеству ФЗ; алгоритмы построения различных покрытий множества ФЗ [12, 15].

Но на практике в проекте РБД присутствуют не только ограничения, соответствующие ФЗ, но и практически все ограничения, приведённые в Таблице 1. В будущих работах планируется создание формальной системы, учитывающей все типы ЭСА, заданные в Таблице 1 и таким образом учитывающие все ограничения, характерные для типичных ПрО. Ниже рассматривается более простой случай, когда исходное множество ЭСА, на основе которого стоится схема РБД, состоит только из связей типа 1..M:0..1. Данный случай похож на классический, когда проект РБД строится только на основе множества ФЗ [12, 15] (связи типа 1..M:1..1 или 1..1:1..1).

Дадим более строгое определение ЭСА типа 1..M:0..1.

Определение 3. Пусть R – схема отношения *r*, $\boldsymbol{A}$ и $\boldsymbol{B}\in R$, пусть задана связь *RS*: $\boldsymbol{A} \xleftrightarrow{1..M:0..1} \boldsymbol{B}$. Тогда отношение *r* удовлетворяет связи типа 1..M:0..1, если $\pi_B(\sigma_{A=a}(r))$ содержит либо *null-total* кортежи, либо строго один *value-total* кортеж, где π – операция проекции, σ – операция селекции.

Далее представим множество правил вывода (аксиом) для ЭСА типа 1..M:0..1. Пусть R – схема отношения и $\boldsymbol{A}$, $\boldsymbol{B}$, $\boldsymbol{C}$ – непересекающиеся подмножества атрибутов схемы R. Тогда для любого отношения *r* со схемой R справедливы следующие правила вывода:

Rule 1. *Пополнение*: $\boldsymbol{A} \xleftrightarrow{1..M:0..1} \boldsymbol{B}$ влечёт за собой $\boldsymbol{AC} \xleftrightarrow{1..M:0..1} \boldsymbol{B}$.

Rule 2. *Проективность*: $\boldsymbol{A} \xleftrightarrow{1..M:0..1} \boldsymbol{BC}$ влечёт за собой $\boldsymbol{A} \xleftrightarrow{1..M:0..1} \boldsymbol{B}$ и $\boldsymbol{A} \xleftrightarrow{1..M:0..1} \boldsymbol{C}$.

Rule 3. *Транзитивность*: $\boldsymbol{A} \xleftrightarrow{1..M:0..1} \boldsymbol{B}$ и $\boldsymbol{B} \xleftrightarrow{1..M:0..1} \boldsymbol{C}$ влечёт за собой $\boldsymbol{A} \xleftrightarrow{1..M:0..1} \boldsymbol{C}$.

С помощью этих правил можно построить замыкание связей типа 1..M:0..1. Доказательство надёжности и полноты системы правил *Rule* 1, *Rule* 2, *Rule* 3 в статье не приводится.

3.2 Алгоритм построения замыкания атрибутов относительно множества связей типа 1..M:0..1

Пусть $\boldsymbol{RS}$ – исходное множество связей типа 1..M:0..1. Алгоритм 1 определяет замыкание атрибутов $\boldsymbol{A}^+$ над $\boldsymbol{A}$ относительно $\boldsymbol{RS}$.

Алгоритм 1 Алгоритм вычисления замыкания атрибутов относительно множества связей типа 1..M:0..1

```
CLOSURE(RS,A)
ВХОД:
1. RS – исходное множество связей типа 1..M:0..1, заданное на схеме U;
2. A – множество атрибутов A⊆U.
ВЫХОД: A+ – замыкание атрибутов над A.
МЕТОД:
CLOSURE := ∅;
OLDCLOSURE := ∅;
BEGIN
 OLDCLOSURE := CLOSURE
 FOR ALL X ←1..M:0..1→ Y∈RS
 DO IF X⊆A ∪ CLOSURE
 THEN CLOSURE := CLOSURE∪Y
END
WHILE(CLOSURE ≠ OLDCLOSURE)
```

Докажем корректность работы Алгоритма 1.

Теорема 1. Алгоритм 1 корректно вычисляет замыкание атрибутов $\boldsymbol{A}^+$ над $\boldsymbol{A}$ относительно некоторого множества связей типа 1..M:0..1 $\boldsymbol{RS}$.

Доказательство. Пусть CLOSURE – множество атрибутов, возвращаемых алгоритмом 1. Нужно доказать эквивалентность CLOSURE и $\boldsymbol{A}^+$. Это можно достичь, доказав, что CLOSURE$\subseteq\boldsymbol{A}^+$ и $\boldsymbol{A}^+\subseteq$CLOSURE.

1. Докажем, что CLOSURE$\subseteq\boldsymbol{A}^+$. Для этого покажем, что если некоторый атрибут $W\in$CLOSURE, то $W\in\boldsymbol{A}^+$. Возможны два случая:

а) Пусть $W\in$CLOSURE и в исходном множестве связей $\boldsymbol{RS}$ существует связь $\boldsymbol{X} \xleftrightarrow{1..M:0..1} \boldsymbol{Y}$, такая, что $\boldsymbol{X}\subseteq\boldsymbol{A}$ и атрибут $W\in\boldsymbol{Y}$. Тогда из связи $\boldsymbol{X} \xleftrightarrow{1..M:0..1} \boldsymbol{Y}$ можно получить связь $\boldsymbol{A} \xleftrightarrow{1..M:0..1} W$, применив правила: *Rule* 1 (*пополнение)* и *Rule* 2 (*проективность)*, следовательно, $W\in\boldsymbol{A}^+$.

б) Пусть $W\in$CLOSURE и в исходном множестве связей $\boldsymbol{RS}$ существуют связь $\boldsymbol{X} \xleftrightarrow{1..M:0..1} \boldsymbol{Y}$, такая, что $\boldsymbol{X}\subseteq\boldsymbol{A}$, и связь $\boldsymbol{Y} \xleftrightarrow{1..M:0..1} \boldsymbol{Z}$, такая, что $W\in\boldsymbol{Z}$. Из этих связей можно получить связь $\boldsymbol{A} \xleftrightarrow{1..M:0..1} W$, применив правила *Rule* 1 (*пополнение)*, *Rule* 2 (*проективность)*, *Rule* 3 (*транзитивность)*, следовательно, $W\in\boldsymbol{A}^+$. Для данного случая возможен вариант, когда между описанными исходными двумя связями будет цепочка из *n* связей, таких, что к каждой связи из этой цепочки можно применить правило *Rule 3* (*транзитивность)*. В этом случае справедливость вывода может быть легко доказана индукцией по числу шагов в цепочке вывода.

2. Докажем, что $\boldsymbol{A}^+\subseteq$CLOSURE. Множество связей, которое можно получить применением правил *Rule* 1, *Rule* 2, *Rule* 3 к исходному множеству связей $\boldsymbol{RS}$, будем называть замыканием связей над $\boldsymbol{RS}$ и обозначать $\boldsymbol{RS}^+$. Пусть некоторый атрибут $Y\in\boldsymbol{A}^+$. Это означает, что в $\boldsymbol{RS}^+$ существует связь $\boldsymbol{A} \xleftrightarrow{1..M:0..1} Y$ и, следовательно, связь $\boldsymbol{A} \xleftrightarrow{1..M:0..1}$

Y может быть получена применением правил *Rule* 1, *Rule* 2, *Rule* 3. Покажем, что если некоторое множество атрибутов $\boldsymbol{Y}\in\boldsymbol{A}^+$, то $\boldsymbol{Y}\in$CLOSURE$^{(i)}$, где CLOSURE$^{(i)}$ – состояние переменной CLOSURE на *i*-м шаге Алгоритма 1. Вышесказанное докажем по индукции. Используем индукцию по числу шагов в цепочке вывода, выдаваемой Алгоритмом 1. Гипотеза индукции: «Если $\boldsymbol{X}\subseteq\boldsymbol{A}$ и по правилам *Rule* 1, *Rule* 2, *Rule* 3 связь $\boldsymbol{X}\xleftrightarrow{1..M:0..1}\boldsymbol{Y}$ следует из $\boldsymbol{RS}$ за не более чем *s* шагов, то каждый атрибут из $\boldsymbol{Y}$ содержится в CLOSURE$^{(s)}$».

Базис: *s*=1. В $\boldsymbol{RS}$ существует некоторая связь $\boldsymbol{X}\xleftrightarrow{1..M:0..1}\boldsymbol{Y}$, тогда каждый атрибут из $\boldsymbol{Y}$ будет тривиально принадлежать CLOSURE$^{(1)}$.

Индукция: *s*>1. В соответствии с гипотезой индукции для некоторой связи $\boldsymbol{V}\xleftrightarrow{1..M:0..1}\boldsymbol{W}$, полученной на шаге вывода *s*–1, $\boldsymbol{W}\subseteq$CLOSURE$^{(S-1)}$. Докажем, что на шаге вывода *s* для связи $\boldsymbol{X}\xleftrightarrow{1..M:0..1}\boldsymbol{Y}$, которая следует из $\boldsymbol{V}\xleftrightarrow{1..M:0..1}\boldsymbol{W}$ по одному из трёх правил вывода (*Rule* 1 пункт «а» доказательства, *Rule* 2 пункт «б», *Rule* 3 пункт «в»), $\boldsymbol{Y}$ будет принадлежать CLOSURE$^{(s)}$.

а) *Rule* 1 (*пополнение).* Пусть существует связь $\boldsymbol{V}\xleftrightarrow{1..M:0..1}\boldsymbol{Y}$, выведенная за менее чем *s* шагов, тогда в соответствии с гипотезой индукции $\boldsymbol{Y}\subseteq$CLOSURE$^{(s-1)}$. Пусть существует множество атрибутов $\boldsymbol{T}$, такое, что $\boldsymbol{VT}\subseteq\boldsymbol{X}\subseteq\boldsymbol{A}$. Применим правило *Rule* 1 и получим связь $\boldsymbol{X}\xleftrightarrow{1..M:0..1}\boldsymbol{Y}$, для которой множество атрибутов $\boldsymbol{Y}$ будет по-прежнему принадлежать CLOSURE. Поэтому на шаге *s* имеем $\boldsymbol{Y}\subseteq$CLOSURE$^{(s)}$.

б) *Rule* 2 (*проективность*). Пусть существует связь $\boldsymbol{X}\xleftrightarrow{1..M:0..1}\boldsymbol{W}$, выведенная за менее чем *s* шагов, тогда в соответствии с гипотезой индукции $\boldsymbol{W}\subseteq$CLOSURE$^{(s-1)}$. Пусть $\boldsymbol{Y}\subseteq\boldsymbol{W}$, применим к $\boldsymbol{X}\xleftrightarrow{1..M:0..1}\boldsymbol{W}$ правило *Rule 2* и получим $\boldsymbol{X}\xleftrightarrow{1..M:0..1}\boldsymbol{Y}$**.** Так как $\boldsymbol{W}\subseteq$CLOSURE и $\boldsymbol{Y}\subseteq\boldsymbol{W}$, то $\boldsymbol{Y}\subseteq$CLOSURE$^{(s)}$.

в) *Rule* 3 (*транзитивность).* Пусть связь $\boldsymbol{X}\xleftrightarrow{1..M:0..1}\boldsymbol{Y}$ следует по правилу *Rule* 3 из предыдущих двух связей $\boldsymbol{X}\xleftrightarrow{1..M:0..1}\boldsymbol{Z}$ и $\boldsymbol{Z}\xleftrightarrow{1..M:0..1}\boldsymbol{Y}$. Связь $\boldsymbol{X}\xleftrightarrow{1..M:0..1}\boldsymbol{Z}$ выводится за *n* шагов ($n < s$), тогда по гипотезе индукции $\boldsymbol{Z}\subseteq$CLOSURE$^{(n)}$. Рассмотрим исполнение алгоритма 1 с $\boldsymbol{Z}$ вместо $\boldsymbol{X}$. В соответствии с гипотезой индукции за *k* шагов можно получить связь $\boldsymbol{Z}\xleftrightarrow{1..M:0..1}\boldsymbol{Y}$, такую, что $\boldsymbol{Y}\subseteq$CLOSURE$^{(k)}$. Учтём следующее Наблюдение 1: пусть $\boldsymbol{RS}$ – множество связей типа 1..M:0..1, $\boldsymbol{A}$ и $\boldsymbol{B}$ – множества атрибутов. Если $\boldsymbol{A}\subseteq\boldsymbol{B}$, то CLOSURE($\boldsymbol{RS}$,$\boldsymbol{A}$)$\subseteq$CLOSURE($\boldsymbol{RS}$,$\boldsymbol{B}$). Выше было получено $\boldsymbol{Z}\subseteq$CLOSURE$^{(n)}$. В соответствии с наблюдением 1, если $\boldsymbol{Z}\subseteq$CLOSURE$^{(n)}$, то CLOSURE$^{(k)}\subseteq$CLOSURE$^{(n+k)}$. Так как множество атрибутов $\boldsymbol{Y}$ содержалось в CLOSURE$^{(k)}$, то $\boldsymbol{Y}$ содержится и в CLOSURE$^{(n+k)}$.

Доказав CLOSURE$\subseteq\boldsymbol{A}^+$ и $\boldsymbol{A}^+\subseteq$CLOSURE, получим, что CLOSURE и $\boldsymbol{A}^+$ эквиваленты и алгоритм 1 корректно вычисляет $\boldsymbol{A}^+$ над $\boldsymbol{A}$ относительно множества связей $\boldsymbol{RS}$.

3.3 Membership-алгоритм для связей типа 1..M:0..1

Пусть $\boldsymbol{RS}$ – исходное множество связей типа 1..M:0..1 и имеется произвольная связь RS_1: $\boldsymbol{A}\xleftrightarrow{M..1:0..1}\boldsymbol{B}$. Если $\boldsymbol{B}\subseteq\boldsymbol{A}^+$, то связь RS_1 принадлежит $\boldsymbol{RS}^+$, то есть может быть выведена из $\boldsymbol{RS}$ через применение правил *Rule* 1, *Rule* 2, *Rule* 3. Принадлежность произвольной связи RS_1 замыканию связей $\boldsymbol{RS}^+$ определяется с помощью Алгоритма 2. Таким образом, Алгоритм 2 выполняет роль membership-алгоритма.

Алгоритм 2 Алгоритм проверки принадлежности произвольной связи типа 1..M:0..1 замыканию множества связей типа 1..M:0..1

```
MEMBER(RS, RS1)

ВХОД:

1. RS – исходное множество связей типа
1..M:0..1;

2. произвольная связь RS1: A ←M..1:0..1→ B.

ВЫХОД: принадлежность связи RS1 замыканию
связей RS+

МЕТОД:

BEGIN

IF B⊆CLOSURE(A,RS) THEN
 RETURN(true)
ELSE
 RETURN(false)
END
```

Решение задачи принадлежности (membership) произвольной связи типа 1..M:0..1 замыканию множества связей типа 1..M:0..1 – $\boldsymbol{RS}^+$ необходимо для построения минимальных покрытий, а это, в свою очередь, необходимо для построения схем баз данных.

4 Заключение

Предложены концепции «элементарная связь атрибутов» и «тип элементарной связи атрибутов». Рассмотрен частный случай, когда множество ЭСА состоит только из связей типа 1..M:0..1. Для этого случая представлены система правил вывода и алгоритм вычисления замыкания атрибутов относительно множества связей типа 1..M:0..1 и алгоритм, позволяющий определить принадлежность произвольной связи типа 1..M:0..1 замыканию связей типа 1..M:0..1. Решение этих задач необходимо для построения минимальных покрытий и дальнейшего построения схем баз данных. Приведено доказательство корректности работы алгоритма вычисления замыкания атрибутов

относительно множества связей типа 1..M:0..1.

При дальнейшем развитии предлагаемого подхода планируется обосновать множество правил вывода (аксиом) для других типов ЭСА, приведённых в Таблице 1. Планируется строить непротиворечивые (conflict-free) множества ЭСА различных типов, обосновать правила вывода при совместном использовании ЭСА различных типов и построить алгоритмы замыкания атрибутов и membership-алгоритмы для этих случаев. Далее, на основе разработанных правил и алгоритмов планируется разработать алгоритмы синтеза и декомпозиции для построения схем РБД. Далее, учитывая разработанную теоретическую базу, планируется разработать прикладное программное обеспечение для поддержки процесса построения и сопровождения проекта РБД.

Литература

[1] Bahmani, A., Naghibzadeh, M., Bahmani B.: Automatic Database Normalization and Primary Key Generation. Electrical and Computer Engineering, 2008. CCECE 2008. Canadian Conference on, pp. 11-16. IEEE, Ontario (2008)

[2] Date, C.J., Darwen, H.: All For One, One For All (2006), Available at: http://www.dcs.warwick.ac.uk/~hugh/TTM/AllforOne.pdf

[3] De Marchi, F., Lopes, S., Petit, J.M., Toumani, F.: Analysis of Existing Databases at the Logical Level: the DBA Companion Project. ACM Sigmod Record, 32 (1), pp. 47-52 (2003)

[4] Dhabe, P.S., Patwardhan, M.S., Deshpande, A.A., Dhore, M.L., Barbadekar, B.V., Abhyankar, H.K.: Articulated Entity Relationship (AER) Diagram for Complete Automation of Relational Database Normalization. Int. J. of Database Management Systems (IJDMS), 2 (2), pp. 84-100 (2010)

[5] Du, H., Wery, L.: Micro: A Normalization Tool for Relational Database Designers. J. of Network and Computer Application, 22 (4), pp. 215-232 (1999)

[6] Halpin, T.: Conceptual Schema and Relation Database Design. 2th ed. Prentice-Hall of Australia Pty., Ltd (1995)

[7] Halpin, T., Morgan, T.: Information Modeling and Relational Databases. 2th ed. Kaufmann Publishers (2008)

[8] Hartmann, S., Link, S.: The Implication Problem of Data Dependencies over SQL Table Definitions: Axiomatic, Algorithmic And Logical Characterizations. ACM Transactions on Database Systems (TODS), 37 (2), pp. 13 (2012)

[9] Kolp, M., Zimanyi, E.: A Relational Database Design Using an ER Approach and Prolog. Proc. of Conf. on Information Systems and Management of Data, Bombay (1995)

[10] Levene, M., Loizou, G.: Axiomatisation of Functional Dependencies in Incomplete Relations. Theoretical Computer Science, 206 (1), pp. 283-300 (1998)

[11] Lovrencich, A., Cubrilo, M., Kishasondi, T.: Modelling Functional Dependencies in Databases using Mathematical Logic. Proc. of the 11th Int. Conf. on Intelligent Engineering Systems. IEEE, Budapest (2007)

[12] Maier, D.: The Theory of Relational Databases. Computer Science Press, Inc. (1983)

[13] Mitrovic, A.: NORMIT: a Web-Enabled Tutor for Database Normalization. Proc. of the Int. Conf. on Computers in Education (ICCE). IEEE, Auckland (2002)

[14] Patwardhan, M.S., Dhabe, P.S., Deshpande, A.A., Londhe, S.G., Dhore, M.L., Abhyankar, H.K.: Diagrammatic Approach for Complete Automation of Relational Database Normalization at Conceptual Level. Int. J. of Database Management Systems (IJDMS), 2 (4), pp. 132-151 (2010)

[15] Ullman, J.: Principles of Database Systems. 2th ed. Computer Science Press, Rockville (1982)

[16] Yazici, A., Ziya, K.: JMathNorm: A Database Normalization Tool using Mathematica. Int. Conf. on Computational Science 2007 (ICCS 2007). Beijing (2007)

[17] Zaniolo, C.: Database Relations with Null Values. Proc. of the 1st ACM SIGACT-SIGMOD Symposium on Principles of Database Systems, pp. 27-33 (1982)

Ключевой доклад 2

Keynote Talk 2

The Astronomical Data Deluge: the Template Case of Photometric Redshifts

(Extended Abstract)

© Giuseppe Longo[1,2] © Massimo Brescia[2,1] © Stefano Cavuoti[1,2]

[1] Department of Physics, University Federico II, Napoli, Italy
[2] INAF Astronomical Observatory of Capodimonte, Napoli, Italy

longo@na.infn.it brescia@oacn.inaf.it cavuoti@na.infn.it

Abstract. Machine learning methods have become crucial to many aspects of astrophysics and cosmology. We focus on the evaluation of photometric redshifts as a template case of classification/regression problem in astronomical data mining. We discuss the general aspects of the problem and some recent work which tries to solve the issues posed by optimal feature selection, missing data and by the evaluation of probability distribution functions.

Keywords: data intensive domains, astrophysics big data, machine learning, photometric redshifts.

1 Introduction

Multiband, multi-epoch digital sky surveys are producing a tsunami of complex, high quality data, which is changing the landscape of astrophysical research. New generation survey telescopes such as the Large Synoptic Survey Telescope (LSST) and Euclid in the optical domain, or the Square Kilometer array (SKA) in the radio domain, will soon produce many tens of TB of processed data every day, and on the long term will provide hundreds of measured parameters for billions of sources. An unprecedented wealth of high quality, accurate and complex data – stored in distributed data centers - that on the long term is expected to revolutionize our understanding of the universe. In order to cope with this data overabundance, all steps of the data understanding chain – acquisition, reduction, analysis, visualization and interpretation – are being deeply transformed and machine learning methods (ML) are becoming crucial at every stage of the process. In particular, modern precision cosmology requires accurate information on both type and redshift (i.e. the distance) for very large (in the hundreds of millions) samples of galaxies. This task cannot be accomplished by means of traditional spectroscopic techniques and in recent years there has been an explosion of alternative methods based on the exploitation of the information contained in multiband photometry: the so called photometric redshifts (hereafter photo-z). A very effective and promising approach to the evaluation of photo-z relies on ML methods. Many different implementation have appeared in the specialized literature based on different flavors of (Multi Layer Perceptrons) MLP's [cf. 1,2,3], random forest [4], nearest neighbors [5], active learning [6], etc. all with their slight advantages and disadvantages.

Proceedings of the XIX International Conference "Data Analytics and Management in Data Intensive Domains" (DAMDID/RCDL'2017), Moscow, Russia, October 10–13, 2017

Therefore, rather than focusing on a specific method, we shall discuss the general aspects of the problems and some ongoing work addressing the main issues: characterization of the knowledge base, feature extraction and selection, missing data and evaluation of errors.

2 Photo-z with ML Methods

2.1 The Knowledge Base

From a ML point of view, the evaluation of photo-z is a classification/regression problem, where the chosen method learns how to estimate the redshift of a galaxy interpolating the knowledge available for a small but significant subsample of objects with known spectroscopic redshifts (knowledge base or KB). After training, the methods (and the underlying mapping function) can be applied to those objects for which the spectroscopic redshift is not available. Data augmentation techniques have been tested but did not lead to reliable results. More promising seems to be the combination of machine learning methods with other techniques, (such as, for instance, template fitting [7]).
This process has two implications, one rather obvious and the other much less so. First, the methods cannot be applied to objects outside of the parameter space sampled by the KB (for instance, fainter than the spectroscopic limit). Second, methods often fail to capture the properties of objects which, being intrinsically rare or peculiar, are not well represented in the KB. Given the complexity of the extragalactic zoo that spans over a very wide variety of observed and physical properties, understanding the properties of the KB becomes crucial. This will be particularly relevant if we take into account that almost all we know about systematic in photo-z comes from optically selected samples, while some surveys of the future will deal with radio (e.g. SKA) or X-ray (e.g. e-Rosita) selected samples. Some recent attempts have been made which are worth mentioning. In

[8] a SOM was used to map the photometric space expected for the Euclid space mission in order also to define the optimal strategy to build the KB.

2.2 Features Extraction and Feature Selection

Digital surveys produce for each observed object many hundreds of parameters that are often highly correlated. These features (i.e. fluxes within a given aperture, radii, concentration indexes, etc.) are usually derived using recipes based on the expertise of astronomers. A pioneering work [9] based on a purely data driven approach, has recently shown that traditional features, almost always fail to capture the subtleties of the information contained in the raw data. This calls for a new way to access the information contained in the astronomical images. While this process is still in its infancy, there are clear signs that deep learning can be greatly beneficial (K. Polsterer, priv. comm.).

In any case, due to both computational constraints and to the need to optimize the dimensionality of the parameter space, feature selection remains a crucial problem that only recently has begun to be properly addressed within the astronomical community. At the moment, two approaches seem to be viable: a brute force approach, where all possible combinations of features are tried until a plateau in the performances (defined by some metrics) is reached [9,10] and Cavuoti (priv. comm.). This approach, however, is computationally demanding and not very flexible. A different path to the identification of the optimal set of features, is currently being implemented by Brescia and collaborators (Brescia et al. 2017, in preparation).

2.3 Missing Data and Non Detection

Most ML methods do not deal effectively with "missing data" (or NAN) and in many cases incomplete data need to be rejected from the sample. This is no longer possible in many modern astronomical applications where incomplete data might affect a quite large fraction of the objects. Furthermore, we need to take into account that in astronomical applications we encounter two types of missing data: "true" missing data (e.g. objects in a region of the sky not observed in a specific band) and "non detection" (e.g. objects which are observed but not detected in one or more photometric band). Dealing with these two types of missing data obviously pose different problems since the latter contain some information (for instance: an upper limit to the flux) that needs to be taken into account. A new approach has been implemented and tested (Cavuoti et al. in preparation) that makes use of a nearest-neighbors approach, to optimize and reconstruct missing information. This approach has been validated on a variety of real data sets.

2.4 Probability Distribution Functions

In many real science applications of photometric redshifts (e.g. weak lensing and shear map reconstruction) one of the main requirements is the need to provide a PDF (Probability Distribution Function) for both the global distribution and the individual objects. Such requirement cannot be met in a trivial way using ML based techniques, since the analytical relation mapping the photometric parameters onto the redshift space is virtually unknown. The tool METAPHOR (*Machine-learning Estimation Tool for Accurate PHOtometric Redshifts*, [11]) was implemented as a modular workflow, whose internal engine for photo-z estimation makes use of MLPQNA (Multi Layer Perceptron with Quasi Newton Approximation; [1]), with the possibility to easily replace the specific machine learning model. METAPHOR takes into account all possible sources of error both internal to the method (e.g. initialization errors) and external (e.g. photometric errors). METAPHOR is independent on the specific ML method used to evaluate the photo-z (it has been extensively tested using several implementation of MLP's and Random Forest algorithm. Recent tests on the KiDS (Kilo Degree survey; [12]) Third Data Release confirmed the robustness of the approach [13].

References

[1] Cavuoti S. et al.: Photometric redshifts with quasi Newton algorithm (MLPQNA). Results in the PHAT1 contest, Astr. & Astroph., 546, 13 (2012)

[2] Brescia M. et al: DAMEWARE: A web cyberinfra-structure for astrophysical data mining, Publ. Astron. Soc. of Pacific, 126, 783 (2014)

[3] Sadeh I. et al., ANNz2: Photometric Redshift and Probability Distribution Function Estimation using Machine Learning, Publ. Astron. Soc. of Pacific, 128 (2016)

[4] Carliles R. et al.: Random Forests for Photometric Redshifts, Astrop. Journ., 712, 511 (2016)

[5] Sheldon E. S. et al.: Photometric Redshift Probability Distributions for Galaxies in the SDSS DR8, Astrop. Journ. Suppl. Series, 2012, 32 (2012)

[6] Bo H. et al.: Active learning applied for photometric redshift estimation of quasars, AAS, 2015IAUGA..2256851H (2015)

[7] Cavuoti S. et al., A cooperative approach among methods for photometric redshifts estimation: an application to KiDS data, MNRAS, 466, 2039 (2017)

[8] Masters D. et al: Mapping the galaxy color-redshift relation: optimal redshift calibration strategies for cosmological surveys, Astrop. Journ., 813, 53 (2015)

[9] Polsterer K. et al., Improving the Performance of Photometric Regression Models via Massive Parallel Feature Selection, proceedings of, Astronomical Data Analysis Software and Systems XXIII, p. 425 (2013)

[10] D'Isanto A. et al: An analysis of feature relevance in the classification of astronomical transients with machine learning methods, MNRAS, 457, 3119-3132 (2016)

[11] Cavuoti S. et al.: METAPHOR: A machine learning based method for the probability density estimation of photometric redshifts, MNRAS 465, 1969 (2016)

[12] de Jong J.T.A. et al: The Third data release of the kilo-degree survey and associated data products, Astr. & Astrop. (arXiv:1703.02991) (2017)

[13] V. Amaro, et al: Machine learning based photometric probability density functions for the KiDS ESO DR3 galaxies, MNRAS (2017)

Проекты анализа данных в астрономии

Data analysis projects in astronomy

Накопление новых знаний о внутреннем устройстве рассеянных звездных скоплений на основе интенсивного использования данных

© С.В. Верещагин © Е.С. Постникова

Институт астрономии Российской академии наук,
Москва, Россия

svvs@ya.ru es_p@list.ru

Аннотация. Предложена методика исследования движений звезд внутри рассеянных звездных скоплений. Она позволяет выявить детали устройства скопления на основе точных измерений астрометрических параметров звезд. На основе последовательного перебора множества скоплений и применения единой методики строится конвейер. Анализ результатов массовой обработки позволит выявить закономерности и связи внутреннего устройства скоплений с их параметрами и положением в Галактике. Актуальности проблеме добавляет то обстоятельство, что запущенный недавно космический телескоп GAIA исследует, в частности, звезды скоплений.

Ключевые слова: звездные каталоги, звездные группы, рассеянные звездные скопления, реестр наименований скоплений, номера звезд, входящих в скопления, AD-диаграмма, апекс звезды.

Accumulation of New Knowledge about the Internal Structure of an Open Star Clusters on the Basis of Intensive Use of Data

© S.V. Vereshchagin © E.S. Postnikova

Institute of Astronomy of Russian academy of Science,
Moscow, Russia

svvs@ya.ru es_p@list.ru

Abstract. A technique for studying the motions of stars inside open star clusters is proposed. It allows revealing the details of the cluster device on the basis of precise measurements of the astrometric parameters of the stars. Successively scanning a lot of clusters and applying a uniform technique, a processing pipeline is built. Analysis of the results of mass processing will reveal the patterns and relationships of the internal arrangement of clusters with their parameters and position in the Galaxy. Actuality of the problem is added by the fact that the recently launched space telescope GAIA investigates, in particular, star clusters.

Keywords: catalogues, star reviews, star groups, open star clusters, register of clusters names, star numbers into clusters, AD-diagram, star apex.

1 Введение

Рассеянные звездные скопления (РЗС) – важные представители населения Млечного Пути. Фундаментальные астрофизические параметры скоплений, такие, как возрасты и массы, могут быть определены надежнее и точнее, чем для отдельных звезд. С их помощью изучают, с одной стороны, образование и начальные стадии эволюции звезд, а с другой – динамическую, фотометрическую и химическую эволюцию Галактики.

В ближайшие годы есть перспектива открыть до ста тысяч скоплений. Учитывая, что в состав скопления входят от десятков до тысяч звезд, становится понятно, что мы имеем дело с большими данными. Этот фактор важен в данной работе, поскольку мы изучаем именно внутреннюю структуру отдельных скоплений. Огромную роль играют уже накопленные знания, без которых использование вновь получаемых данных неэффективно. Это намного увеличивает объем обрабатываемой информации.

Накопления знаний о внутреннем устройстве скоплений имеет как теоретический, так и наблюдательный аспекты. Учитывая широту материала и специфику работы, рассмотрим лишь наблюдательный аспект. В рамках наблюдений

Труды XIX Международной конференции «Аналитика и управление данными в областях с интенсивным использованием данных» (DAMDID/ RCDL'2017), Москва, Россия, 10–13 октября 2017 года

перспективными представляются исследования звездного состава скоплений, обнаружение разного рода особенностей в их строении, а также необычных звезд и экзопланет. Последнее связано с тем, что любое скопление представляет ансамбль звезд различных масс и светимостей, сосредоточенных, как правило, на небольшом участке звездного неба. Это позволяет исследователям одновременно видеть на небольшой площади и изучать звезды с различными свойствами.

Мало изученной с точки зрения наблюдений является внутренняя структура скоплений. Давно известно, что распределение звезд в рассеянных и даже шаровых скоплениях, а также в поле Галактики не является строго однородным, наблюдения показывают звездные сгустки разных масштабов. Их природа пока не до конца выяснена. Перспективное направление – это поиск ранее неизвестных субструктур среди населяющих скопление звезд. Задача обнаружения и каталогизации таких объектов представляет интерес и рассматривается в данной статье. Первичным источником информации здесь являются массовые обзоры звездного неба. По объему они, несомненно, принадлежат к категории больших данных и требуют технологий интенсивного использования.

Мы рассматриваем лишь рассеянные звездные скопления, которые наиболее близки по расстояниям к Солнцу. Шаровые скопления расположены значительно дальше и к ним, как правило, наши методики не применимы. Кроме того, для близких скоплений данные измерений наиболее надежны, что позволяет рассчитывать на достоверные результаты и изучать детали их внутренней структуры. Близкие скопления служат для изучения многих аспектов, включая формирование звезд, звездные структуры, звездное разнообразие и околозвездные процессы, в том числе и формирование планет [25].

Цель работы – построение конвейерной системы отбора и обработки данных о потенциально интересных объектах. Это позволит получить новые знания о строении рассеянных звездных скоплений, эволюции Галактики и ее подсистем. Наилучший вариант первичного рассмотрения с точки зрения надежности наблюдений – это ближайшие к Солнцу звездные потоки и скопления с достаточным набором данных об их звездном составе. Скопления, для которых составлены каталоги с определением необходимого набора параметров, также рассматриваются.

Есть различные методы изучения звездных скоплений, например, метод «движущегося скопления» [18], [12]. Этот метод хорошо подходит до расстояний от Солнца в несколько сотен парсек. Метод тригонометрических параллаксов применим также к достаточно близким скоплениям [20]. Известны методы, опирающиеся на химический состав предполагаемых звезд скопления [16], Интересны методы автоматического и визуального поиска флуктуаций звездной плотности по астрометрическим и фотометрическим данным [6], [10].

Как уже понятно, изучение звездных скоплений мы ведем, отталкиваясь не только от задач, но и от данных, рост которых в последние годы обещает быть очень большим. Особенно это актуально после публикации в открытом доступе первых результатов телескопа GAIA (Global Astrometric Interferometer for Astrophysics) [11].

Центральное место в нашем исследовании занимает метод АД диаграмм, основанный на поиске закономерностей в пространственных движениях звезд внутри скоплений. Он позволяет изучать внутреннюю структуру скоплений и выявлять различные закономерности в движениях звезд членов скопления. Метод был апробирован на многих объектах и требует более тщательной алгоритмизации с использованием средств работы с большими данными.

Структура статьи такова. Во втором разделе дано представление об объекте исследования, в третьем разделе приведены названия публичных архивов данных. Четвертый раздел посвящен динамике увеличения количества данных о скоплениях. В пятом разделе рассмотрены апробированная методика и перспективы выстраивания обработки в последовательность однотипных действий – конвейер, направленный на применение к среде больших данных о скоплениях. В шестом разделе представлены обсуждение и выводы.

2 Рассеянные звездные скопления

2.1 Современное состояние исследований

Звездные скопления традиционно изучаются либо как подсистема объектов Галактики, либо как гравитирующая звездная система. Детали внутреннего устройства РЗС, такие, как группы звезд, изучаются не столь давно и представляют собой интерес. Понятие «группа звезд» появилось, по крайней мере, в 1969 году [34]. Вопрос существования пространственных и кинематических групп звезд в коронах РЗС пока малоизучен, хотя набирает популярность в связи с ожиданием астрометрических результатов миссии GAIA [19]. Об истории вопроса см. [37].

Интересна и другая сторона медали – найдены ли какие-либо физические процессы, которые приведут к образованию групп? Существует несколько подходов, каждый из которых проливает свет на эту проблему.

Впервые ван Альбада в 1968-м году в [29] показал возможность формирования широких двойных и кратных звезд с характерными размерами примерно 10 а.е. и более. Такое событие может происходить в результате распада

небольших звездных групп с размерами от 10**2 до 10**5 а.е. Для справки добавим, что 1 а.е. составляет 149 597 871 км и 1 пк = 206264.8 а.е. Максимальный размер получается равным примерно 0.5 пк.

Об образовании звездных систем высокого порядка кратности в звездных скоплениях: в обзоре Ларсона [15] показано, что большинство звезд в окрестности Солнца (одиночных, двойных и устойчивых кратных) могло образоваться в результате распада неиерархических малых групп звезд, содержащих от нескольких штук до нескольких десятков объектов. Процесс образования кратных систем в скоплениях рассмотрен в [30].

Существуют группы более высокого порядка, состоящие из скоплений, они обнаружены в [9]. Выделены четыре структуры, видимые в вэйвлет пространстве, которые соответствуют потоку Геркулеса, Плеядам и Гиадам, группе Сириуса. Соответствующие динамические модели свидетельствуют об их резонансном происхождении в точках Лагранжа балджа Галактики, от спиральных ветвей или комбинации этих воздействий. Исследование [9] подтверждает резонансную природу происхождения потоков Плеяд, Гиад и Большой Медведицы (Сириуса).

2.2 Поиски неизвестных скоплений

Картина неба, видимая глазом, обманчива – мы легко находим Плеяды, но не видим многие тысячи подобных скоплений, которые либо малы, либо расположены достаточно далеко от Солнца и сливаются со звездным фоном. На небесной сфере условно можно выделить две звездные фракции – фон и флуктуации звездной плотности. На Рис. 1 показаны РЗС в созвездии Персея – NGC 869 и NGC 884 (New General Catalogue of Nebulae and Clusters of Stars), известные также как двойное скопление в Персее. Для наблюдателя они выглядят как флуктуации звездной плотности. Поиск подобных флуктуаций является основой, лежащей в открытии ранее неизвестных скоплений.

Флуктуации могут иметь генетическую природу или носить случайный характер. Искусство ученого заключается не только в том, чтобы суметь ее найти, но и определить, что найдена именно физически связанная группировка. Для этого применяются критерии, основанные на общности собственных движений звезд на небе. Дополнительные критерии основаны на определении и использовании возрастов и химических составов звезд, а также их фотометрических характеристик. Вся совокупность перечисленной информации может свидетельствовать о том, что перед нами физически связанная группа – звездное скопление.

Вид скопления на небе определяется двумя факторами – расстоянием от Солнца и физическими размерами. Многие скопления без дополнительных тестов затруднительно отличить от случайных флуктуаций плотности. Так, крупное скопление, расположенное близко к наблюдателю, может занимать настолько большую площадь на поверхности неба, что будет неотличимо от фона.

Часто необходимо понять, что обнаружено новое, ранее неизвестное скопление. Для этого нужно провести отождествление найденной группы звезд по каталогам с данными об известных скоплениях. Если среди известных скоплений ее нет, то выдвигается предположение о том, что обнаружено новое скопление. Более детально эти вопросы рассмотрены в [22], где опубликованы результаты поиска неизвестных скоплений по видимым уплотнениям на небе.

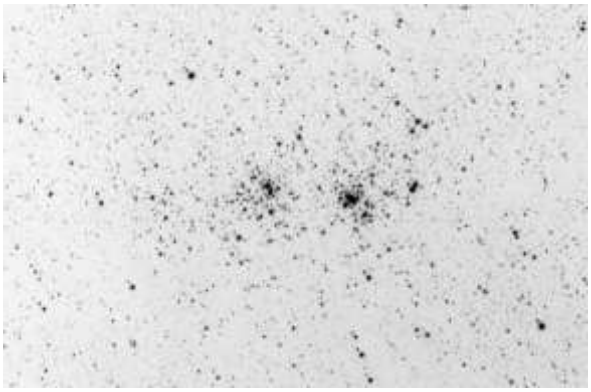


Рисунок 1 Так выглядят рассеянные звездные скопления в небольшой телескоп. Фрагмент фотопластинки с изображением скопления h and χ Персея. Фотопластинка получена на 40-см Астрографе Звенигородской обсерватории. Наблюдатель В.П. Осипенко

В [23] представлен результат поиска неизвестных скоплений. Использован анализ пространства скоростей, где звезды обнаруженной группировки выделяются совместным движением в Галактике.

3 Архивы данных по скоплениям

Существует множество баз данных (БД), где можно найти каталоги скоплений, например, VizieR [31], SIMBAD (Set of Identifications, Measurements and Bibliography for Astronomical Data) [26]. Есть БД, содержащие публикации статей, такие, как ADS NASA (Astrophysics Data System) [27], ScienceDirect (данные издательства Elsevier) [24], IOPscIence [13], Wiley Online Library [33]. Самая крупная – ADS NASA – включает информацию более чем о 7 млн. документов. Упомянем еще БД и каталоги – это WEBDA (A site Devoted to Stellar Clusters in the Galaxy and the Magellanic Clouds) [32], Линга [17], Альтер и Рупрехт [1], Бархатова [35], Пискунов 0, которые имеют в основном исторический интерес, хотя в рамках нашей задачи информация из них актуальна для сравнения результатов. Наименования скоплений в разных каталогах различны, что создает проблемы для

отождествления. Эти наименования представляют метаданные, соответствие между которыми имеются, например, в SIMBAD [26].

Не все из перечисленных БД и каталогов пополняются новыми данными. БД Диас [5] представляет собой приятное исключение и продолжает пополняться новыми данными. Она ценна также тем, что в ней использованы активные ссылки (в каждой строке) на WEBDA [32], Линга [17], Харченко и др. [14]. Также эта БД включает библиотечные коды публикаций.

Коллективом авторов (список участников см. в [14]) создается Глобальный обзор звездных скоплений Млечного Пути (Milky Way Global Survey of Star Clusters, MWSC). Он включает десять каталогов и содержит данные обо всех скоплениях, известных к настоящему времени. Система каталогов MWSC является наиболее полной и часто цитируемой в современных исследованиях скоплений. Основные результаты приведены в [14] – это каталог, содержащий все значения метаданных скоплений. По астрометрии это: экваториальные координаты, диаметр скопления, собственные движения и лучевые скорости звезд, их возрасты, покраснения, расстояния от Солнца, приливные радиусы. Активными в каталоге [14] для каждого из скоплений являются дополнительные страницы со звездными картами и диаграммами цвет–величина. Эти диаграммы позволяют узнать звездный состав и возраст скопления. Сейчас к MWSC активно подключаются данные GAIA [11].

4 Рост данных, проект GAIA

В последнее время в астрофизике и других научных дисциплинах все более популярным становится направление `”data science” – извлечение научных фактов из больших массивов данных. Увеличение их объема идет за счет роста числа новых открытий и роста информации за счет новых наблюдений и публикаций параметров о каждом из известных скоплений.

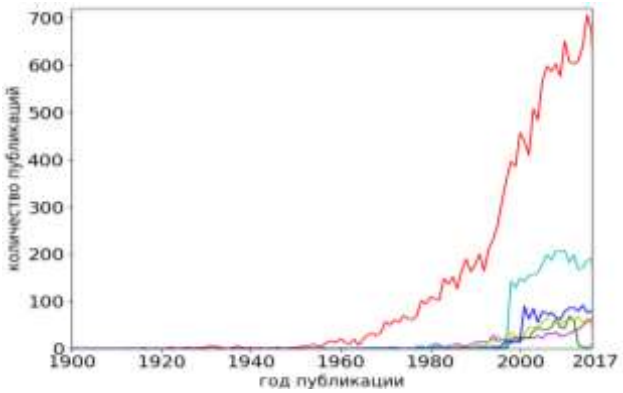


Рисунок 2 Результат поиска статей по ключевым словам “open&cluster”. Верхняя кривая – для ADS (всего 15261 публикаций). В середине – IOPscIence (3489). Менее максимума в сто публикаций – Scince Direct (858), Wiley Online Library (625), VizeR (1152), A&A [2] (1365)

Разнообразие состава и рост объема информации (измеренной в публикациях) о скоплениях показано на Рис. 2. Резкий скачок числа публикаций начался, как видим на Рис. 2, с 2000-х годов. Это понятно – ведь рост данных в астрономии связан с введением в строй новых дорогостоящих телескопов. Таким образом, не удивительны скачки роста информации в периоды появления известных телескопов, что мы видим на Рис. 2.

Статистика строительства телескопов такова: рефракторы с диаметром объектива больше 70 см (это 11 телескопов) были построены в период 1880–1917 гг., телескопы-рефлекторы с диаметром зеркала 6 м и более (14 телескопов) – в период 1975–2005 гг. Известный Паломарский 5.1-м телескоп им. Хейла был введен в строй в 1948 г. Именно с этого момента начался заметный рост информации (верхняя кривая на Рис. 2). Появление новых БД (все прочие кривые) началось с появлением больших телескопов. Нынче запущено множество космических аппаратов с телескопами на борту. Особое место занимает проект GAIA. Обзор GAIA [11] включает приблизительно 1 млрд. звезд, что уже сравнимо с населением Галактики и составляет приблизительно 1% ее звездного населения. Предельная звездная величина равна 20 в системе G (на интервале длин волн от 400 до 1000 нм). Микросекундная точность измерений позволяет получить новую информацию о движениях звезд внутри скоплений. Точность достигнута во многом благодаря сверхдальнему (более 1 млн. км от Земли) расположению ИСЗ в точке Лагранжа (L2), исключающей влияние на положение аппарата гравитации от Земли–Луны и Солнца. Исключаются также засветка отраженным светом Солнца от Земли и Луны, а также влияние переходов из света в тень. Такие переходы мешают работе околоземных аппаратов, таких, как телескоп Хаббл. Кроме того, поддержание аппарата в точке L2 энергетически выгодно.

В упомянутом каталоге [14] каталогизировано 3754 РЗС. Это составляет всего 0.3% скоплений галактического диска. В обзоре GAIA степень охвата РЗС составляет около 1%, предоставляя возможность открыть еще 10 тыс. скоплений. В составе Млечного Пути может быть, как минимум, 100 тыс. скоплений. В нашей работе [36] сделана оценка числа скоплений, которые даст проект GAIA в целом. В далеком будущем можно открыть около 1 млн. скоплений, которые расположены в пределах Млечного Пути, состоящем из приблизительно 100 млрд. звезд.

5 Применение АД-диаграмм

Замечено, что существует множество закономерностей движений звезд в Галактике. Так, например, они участвуют во вращении галактического диска. Есть звезды, образующие потоки, и есть «убегающие» звезды, которые

необъяснимо быстро движутся относительно Солнца. Координаты точки пространства, в направлении которой наблюдается движение звезды, называется ее апексом. Рассмотрим метод диаграмм апексов (AD-диаграмм). AD-диаграмма представляет собой распределение апексов звезд в экваториальной системе координат. Координаты звездных апексов получаются из решения геометрической задачи, в которой находятся пересечения векторов пространственных скоростей звезд с небесной сферой, при этом начала векторов перемещены в точку наблюдений. По аналогии с обычным апексом координаты этих точек в экваториальной системе обозначены как A для прямого восхождения и D – для склонения. Их можно назвать индивидуальными апексами звезд. Формальное описание метода, техника построения диаграмм и формулы для определения эллипсов ошибок можно найти в [3]. Отметим, что эллипсы ошибок (важных в любой работе) можно определить только для звезд каталога Hipparcos, в котором имеются необходимые коэффициенты корреляции между астрометрическими параметрами. Этот метод применялся для исследования Гиад, Яслей, скоплений и групп в Орионе и еще некоторых скоплений. Картинка, представляющая собой распределение апексов звезд, позволяет выявить закономерности движений звезд не только внутри скопления, но и в околосолнечном пространстве. Так, на Рис. 3 показано распределение 249603 звезд с наиболее точно измеренными скоростями и расстояниями от Солнца ([21], [8]) на плоской координатной проекции.

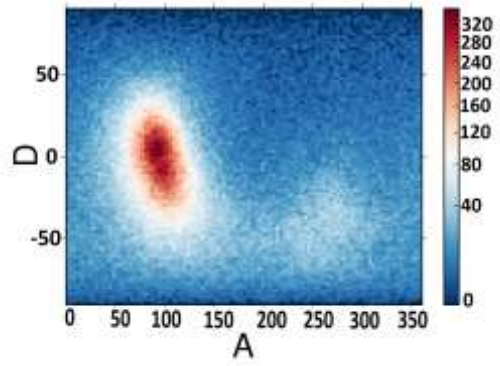


Рисунок 3 АД-диаграмма звезд окрестностей Солнца. Плотности звезд показаны на шкале в правой части рисунка

Наряду с классическими методами исследования наш подход позволяет уделить особое внимание именно внутренней кинематической структуре скоплений.

6 Обсуждение и выводы

6.1 Структура конвейера по накоплению знаний о внутреннем устройстве скоплений

На Рис. 4 представлена схема конвейерной обработки данных по РЗС в рамках предлагаемой парадигмы. Почему такая обработка возникла и стала возможной? Открыто, как уже было отмечено, достаточно много скоплений, для которых возможно применение единой методики обработки методом AD-диаграмм. Скопления различаются по возрастам, числу звезд, диаметрам, степени концентрации звезд к центру и многому другому. Кроме того, они различаются положениями в пространстве относительно Солнца, спиральных ветвей Галактики и т. п. Естественно, что скопления различаются также и диаграммами апексов (в частности, наличием или отсутствием движущихся в них субструктур или групп). Конвейерная обработка позволит накопить информацию об AD-диаграммах множества скоплений. Далее путем сравнительного анализа можно сделать выводы о связи внутренней структуры с другими метаданными. Это в свою очередь позволит получить новые знания и сделать открытия новых закономерностей, касающиеся населения диска Галактики.

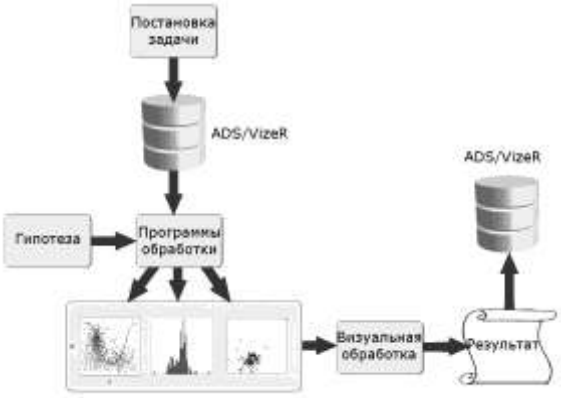


Рисунок 4 Конвейер обработки информации. Включает постановку задачи на основе анализа больших данных (здесь ADS [27] и VizieR [31]) и выдвигаемой гипотезы. В центре – обработка информации по заданным алгоритмам. Полученный ряд изображений проходит визуальную обработку и получение новых результатов. В случае успеха результаты публикуются и в итоге попадают в те же большие данные

На основе отработанных методик проводится анализ устройства рассеянных звездных скоплений. При выборе скоплений учитываются не только физическое многообразие скоплений, но и эффекты, связанные с различиями их положений в пространстве как относительно Солнца, так и внутри Галактики.

Основу обработки составляет метод AD-диаграмм. Он позволяет определить общее направление движения скопления в пространстве и находить возможные внутренние структуры внутри скопления. Накопление результатов о группах в скоплениях осуществляется как в публикациях (в итоге в ADS [27] и VizieR [31]), так и размещением полученных данных в SIMBAD [26].

6.2 Метаданные

По теме и вопросам, затронутым в этой работе, можно выделить метаданные. Это звездные каталоги, звездные группы, рассеянные звездные скопления, реестр наименований скоплений, номера звезд, входящих в скопления, AD-диаграмма, апекс звезды.

6.3 Выводы

Точные позиционные наблюдения представляют собой особую ценность. Они позволяют делать выводы о звездном составе и морфологии скоплений. Данный проект направлен на дальнейшее расширение и улучшение данных о подсистеме звездных скоплений Галактики. Основой являются каталоги нового поколения, которые создаются по наблюдениям космического телескопа GAIA. Конечной целью проекта является накопление информации о строении скоплений в едином формате путем конвейерной обработки данных. Такая информация создаст базу для выявления звездных субструктур внутри скоплений как основы новых знаний о звездных системах. Эти выводы можно распространить и на скопления других галактик. Таким образом, в ближайшее время будут получены большие объемы новых данных, и предлагаемая методика позволит эффективно их обрабатывать для извлечения новых знаний о скоплениях и Галактике в целом.

Звездные потоки в нашей Галактике активно изучаются. Их кинематика и структура проливают свет на детали процесса формирования звездного гало. Тематика проекта очень актуальна.

Создана и применяется система конвейерной обработки данных о скоплениях, Рис. 4. Система включает отбор подходящих скоплений, определение необходимых параметров и выдвижение гипотез. Автоматизированные методы позволяют получать результаты, которые могут содержать новые знания и быть опубликованы. В рамках парадигмы применения диаграмм апексов показан пример открытия субструктур в короне потока БМ [4]. Примеры работы с конвейером – NGC 188, M67 [7], [28]. Удобным хранилищем данных является БД VizieR.

К методическим результатам работы относится разработка процесса исследований индивидуальных скоплений для создания справочного каталога апексов скоплений, а также уникальных научных веб-приложений. С помощью накопленной информации предполагается изучить также аспекты кинематики диска Галактики с выходом на построение его деталей и определение других физических параметров.

Благодарности

Работа частично поддержана Российским фондом фундаментальных исследований (проект 16-52-12027). Е. С. Постникова частично поддержана грантом Президента Российской Федерации для государственной поддержки ведущих научных школ Российской Федерации, грант НШ-9951.2016.2. Мы использовали базу данных SIMBAD, работающую в CDS, Страсбург, Франция. Авторы благодарны рецензентам за ценные замечания и рекомендации.

Литература

[1] Alter, G., Israel, B. Y., Ruprecht, J.: Catalogue of Star Clusters and Associations. Astronomical Institute Czechoslovakia, Prague (1964). doi: 1964csca.book.....A

[2] Astronomy & Astrophysics. Worldwide Astronomical and Astrophysical Research. https://www.aanda. org

[3] Chupina, N.V., Reva, V.G., Vereshchagin, S.V.: The Geometry of Stellar Motions in the Nucleus Region of the Ursa Major Kinematic Group. Astron. Astrophys. 371, pp. 115-122 (2001). doi: 10.1051/0004-6361:20010337

[4] Chupina, N.V., Reva, V.G., Vereshchagin, S.V.: Kinematic Structure of the Corona of the Ursa Major Flow Found Using Proper Motions and Radial Velocities of Single Stars. Astron. Astrophys. 451, pp. 909-916 (2006). doi: 10.1051/0004-6361:20054009

[5] Dias, W.S., Alessi, B.S., Moitinho, A., Lepine, J.R.D.: New Catalog of Optically Visible Open Clusters and Candidates. Astron. Astrophys. 389, pp. 871 (2002). doi: 10.1051/0004-6361:20020668

[6] Drake, A.J.: Cluster Candidates from the USNO-A2.0 Catalogue. Astron. Astrophys., 435 (2), pp. 545-550 (2005). doi: 10.1051/0004-6361:20041568

[7] Elsanhoury, W.H., Haroon, A.A., Chupina, N.V., Vereshchagin, S. V., Sariya, Devesh P., Yadav, R.K.S., Jiang, Ing-Guey: 2MASS Photometry and Kinematical Studies of Open Cluster NGC 188. New Astron., 49, pp. 32-37 (2016). doi: 10.1016/j.newast.2016.06.002

[8] Famaey, B., Jorissen, A., Luri, X., Mayor, M., Udry, S., Dejonghe, H., Turon, C.: Local Kinematics of K and M Giants from CORAVEL/Hipparcos/Tycho-2 Data. Revisiting the Concept of Superclusters. Astron. Astrophys., 430, pp. 165-186 (2005). doi: 10.1051/0004-6361:20041272

[9] Famaey, B., Siebert, A., Jorissen, A.: On the Age Heterogeneity of the Pleiades, Hyades, and Sirius Moving Groups. Astron. Astrophys., 483 (2), pp. 453-459 (2008). doi: 10.1051/0004-6361:20078979

[10] Froebrich, D., Scholz, A., Raftery, C. L.: A Systematic Survey for Infrared Star Clusters with |b| <20° using 2MASS. Mon. Not. R. Astron. Soc., 374, p. 399 (2007). doi: 10.1111/j.1365-2966.2006.11148.x

[11] GAIA DR1 (Gaia Collaboration 2016) A. G. A. Brown e. a.: Gaia Data Release 1 Summary of the Astrometric, Photometric, and Survey Properties. Astron. Astrophys., 595, id.A2, 23, pp. (I/337/tgas) (2016). doi: 10.1051/0004-6361/201629512

[12] Galli, P.A.B., Moraux, E., Bouy, H., Bouvier, J., Olivares, J., Teixeira, R.: A Revised Moving Cluster Distance to the Pleiades Open Cluster. Astron. Astrophys., 598, A48, p. 22 (2017). doi: 10.1051/0004-6361/201629239

[13] IOPscience. http:// http://iopscience.iop.org

[14] Kharchenko, N.V., Piskunov, A.E., Schilbach, E., Röser, S., Scholz, R.-D.: Global Survey of Star Clusters in the Milky Way. II. The Catalogue of Basic Parameters. Astron. Astrophys., 558 (A53), pp. 1-8 (2013). doi: 10.1051/0004-6361/201322302

[15] Larson, R.B.: Implications of Binary Properties for Theories of Star Formation. The Formation of Binary Stars. Proc. of IAU Symp. 200, held 10–15 April 2000, Potsdam, Germany, Ed. H. Zinnecker and R. D. Mathieu, p. 93 (2001)

[16] Liu, F., Yong, D., Asplund, M., Ramírez, I., Meléndez, J.: The Hyades Open Cluster is Chemically Inhomogeneous. Mon. Not. R. Astron. Soc., 457 (4), pp. 3934-3948 (2016). doi:10.1093/mnras/ stw247

[17] Lynga, G.: VizieR On-line Data Catalog. Open Cluster Data. VII/92A. 5th Edition (1987). Originally published in: Lund Observatory (1995)

[18] Mamajek, E.E.: A Moving Cluster Distance to the Exoplanet 2M1207b in the TW Hydrae Association. Astrophys. J., 634 (2), pp. 1385-1394 (2005). doi: 10.1086/468181

[19] Mamajek, E.E.: A Pre-Gaia Census of Nearby Stellar Groups. In: Young Stars & Planets Near the Sun, Proc. of the Int. Astronomical Union, IAU Symposium, 314. pp. 21-26 (2016). doi: 10.1017/ S1743921315006250

[20] Melis, C., Reid, M.J., Mioduszewski, A.J., Stauffer, J.R., Bower, G.C.: A VLBI Resolution of the Pleiades Distance Controversy. Science, 345 (6200), pp. 1029-1032 (2014). doi: 10.1126/science.1256101

[21] Kunder, A., and 53 coauthors: The Radial Velocity Experiment (RAVE): Fifth Data Release. Astron. J., 153 (2), article id. 75, 30 p. (2017). doi: 10.3847/1538-3881/153/2/75

[22] Schmeja, S., Kharchenko, N.V., Piskunov, A.E., Röser, S., Schilbach, E., Froebrich, D., Scholz, R.-D.: Global Survey of Star Clusters in the Milky Way. III. 139 New Open Clusters at high Galactic Latitudes. Astron. Astrophys., 568 (A51), pp. 1-9 (2014). doi: 10.1051/0004-6361/201322720

[23] Scholz, R.-D., Kharchenko, N.V., Piskunov, A.E., Röser, S., Schilbach, E.: Global Survey of Star Clusters in the Milky Way. IV. 63 New Open Clusters Detected by Proper Motions. Astron. Astrophys., 581 (A39), pp. 1-15 (2015). doi: 10.1051/0004-6361/201526312

[24] Scince direct. http://www.sciencedirect.com/

[25] Siegler, N., Muzerolle, J., Young, E.T., Rieke, G.H., Mamajek, E.E., Trilling, D.E., Gorlova, N., Su, K.Y.L.: Spitzer 24 μm Observations of Open Cluster IC 2391 and Debris Disk Evolution of FGK Stars. Astrophys. J., 654, pp. 580-594 (2007). doi: 10.1086/509042

[26] SIMBAD Astronomical Database. http://simbad.u-strasbg.fr/simbad/

[27] The SAO/NASA Astrophysics Data System. http://www.adsabs.harvard.edu

[28] Vereshchagin, S.V., Chupina, N.V., Sariya, D.P., Yadav, R.K.S., Kumar, B.: Apex Determination and Detection of Stellar Clumps in the Open Cluster M 67. New Astron. 31, pp.43-50 (2014). doi: 10.1016/j.newast.2014.02.008

[29] van Albada, T.S.: The Evolution of Small Stellar Systems and its Implications for the Formation of Double Stars. Bull. Astron. Inst. Netherlands, 20, p. 57 (1968)

[30] van den Berk, J., Portegies Zwart, S.F., McMillan, S.L.W.: The Formation of Higher Order Hierarchical Systems in Star Clusters. Mon. Not. R. Astron. Soc., 379 (1), pp. 111-122 (2007). doi: 10.1111/j.1365-2966.2007.11913.x

[31] VizieR Catalogue Service. http://vizier.u-strasbg.fr

[32] WEBDA. A Site Devoted to Stellar Clusters in the Galaxy and the Magellanic Clouds. http://www. univie.ac.at/webda/webda.html

[33] Wiley Online Library. http://onlinelibrary.wiley.com

[34] Walker, M.F.: Studies of Extremely Young Clusters. V. Stars in the Vicinity of the Orion Nebula. Astrophys. J., 155, p. 447 (1969). doi: 10.1086/ 149881

[35] Бархатова, К.А.: Атлас диаграмм цвет – светимость рассеянных звездных скоплений. М.: Изд-во АН СССР, 127 с. (1958)

[36] Верещагин, С.В., Чупина, Н.В., Фионов, А.С.: Звездные скопления: развитие знаний на основе интенсивного использования данных. Труды XVIII Межд. конф. «Аналитика и управление данными в областях с интенсивным использованием данных», DAMDID/ RCDL'2016, под ред. Л.А. Калиниченко, Я. Манолопулоса, С.О. Кузнецова, сс. 323-327 (2016)

[37] Рубинов, А.В., Орлов, В.В.: Задача N тел в звездной динамике. Учебное пособие. Санкт-Петербург: Изд-во С-Пб. университета, 97 с. (2008)

Пискунов, А.Э.: Каталог масс и возрастов 68 рассеянных скоплений. М.: Изд-во «Астрономический Совет АН СССР» (1977)

Короткие транзиентные гамма-события в эксперименте SPI/INTEGRAL: поиск, классификация и интерпретация

© П.Ю. Минаев © А.С. Позаненко

Институт космических исследований Российской академии наук,
Москва, Россия

minaevp@mail.ru apozanen@iki.rssi.ru

Аннотация. Рассмотрены возможности поиска и анализа транзиентных гамма-событий различной природы в архивных данных гамма-спектрометра SPI космической обсерватории INTEGRAL. Обсуждены проблемы обработки массивов наблюдательных данных эксперимента, в том числе алгоритма поиска и методики автоматической классификации обнаруженных событий на основе комплекса критериев. Кратко приведены результаты анализа архивных данных эксперимента SPI/INTEGRAL, полученных за период 2003–2010 гг.

Ключевые слова: космические гамма-всплески, GRB, гамма-всплески земного происхождения, TGF, SGR, AXP, поиск, классификация, каталог, INTEGRAL/SPI.

Short Gamma-ray Transients in SPI/INTEGRAL: Search, Classification and Interpretation

© P.Yu. Minaev © A.S. Pozanenko

Space Research Institute of Russian Academy of sciences,
Moscow, Russia

minaevp@mail.ru apozanen@iki.rssi.ru

Abstract. The possibilities of searching and analyzing gamma-ray transients of various nature in the archival data of the SPI spectrometer of the INTEGRAL space observatory are considered. The problems of processing the arrays of row observational data of the experiment, including the search algorithm and the method of automatic classification of detected events based on a set of criteria are discussed. The results of the analysis of the archived data of the SPI / INTEGRAL experiment obtained for the period 2003–2010 are briefly presented.

Keywords: cosmic gamma-ray bursts, GRB, terrestrial gamma-ray flashes, TGF, SGR, AXP, search, classification, catalog, INTEGRAL/SPI.

1 Введение

Одной из актуальных задач современной астрофизики высоких энергий является исследование гамма-всплесков космического (GRB) и земного (TGF) происхождения. Космические гамма-всплески – одни из самых мощных взрывов во Вселенной – наблюдаются как спорадические вспышки гамма-излучения длительностью 0.1–100 с в энергетическом диапазоне выше 10 кэВ [1]. Гамма-всплески земного происхождения значительно короче (менее 1 мс). Считается, что они генерируются в верхней атмосфере Земли при пробое на убегающих электронах и сопровождают грозовую активность [2, 3].

Большинство современных космических гамма-телескопов (в том числе, спектрометр SPI/INTEGRAL) позволяет регистрировать отдельные гамма-кванты, записывая момент регистрации отсчета, его энергию и некоторые другие параметры (например, номер детектора или пикселя матрицы, в котором произошла регистрация) [1, 4]. Это открывает большие возможности в анализе наблюдательных данных, в том числе разработки алгоритмов поиска транзиентных событий различных типов, а также методики автоматической классификации обнаруженных событий по определенным критериям. Автоматическая классификация событий в рамках анализа значительных массивов данных (более 10 Тб для эксперимента SPI), накопленных за несколько лет наблюдений, имеет чрезвычайно большую роль вследствие огромного числа (более 15000 за год наблюдений) «ложных» срабатываний алгоритма поиска событий,

Труды XIX Международной конференции «Аналитика и управление данными в областях с интенсивным использованием данных» (DAMDID/ RCDL'2017), Москва, Россия, 10–13 октября 2017 года

связанных с различными инструментальными эффектами (например, взаимодействие заряженных частиц с детектором).

В данной работе рассмотрены оригинальные алгоритмы поиска и классификации обнаруженных событий в архивных данных эксперимента SPI/INTEGRAL, а также интерпретация полученных результатов.

2 Эксперимент SPI/INTEGRAL

Обсерватория INTEGRAL была запущена 17 октября 2002 года на высокоэллиптичную орбиту (перигей начальной орбиты – 9 тыс. км, апогей – 153 тыс. км) с периодом 72 часа [5]. На обсерватории размещены два основных гамма-телескопа (IBIS/ISGRI, SPI) и несколько вспомогательных телескопов (JEM-X, OMC, SPI-ACS). Все апертурные телескопы (SPI, IBIS/ISGRI, JEM-X, OMC) соосны, но форма и размер полей зрения различны.

Гамма-спектрометр SPI состоит из 19 детекторов шестиугольной формы, изготовленных из сверхчистого германия, с общей геометрической площадью 508 $см^2$ [6]. Для построения изображений используется кодирующая маска, изготовленная из вольфрама. Спектральное разрешение спектрометра SPI/INTEGRAL достигает значения 2.2 кэВ @ 1.33 МэВ – одно из лучших на момент запуска обсерватории (2002 г). Энергетический диапазон чувствительности 20 кэВ – 8 МэВ. Полное поле зрения телескопа составляет 30^o.

Для увеличения чувствительности телескопа SPI за счет устранения фона, связанного с взаимодействием аппаратуры с космическими лучами, используется антисовпадательная защита SPI-ACS, состоящая из 91 кристалла германата висмута (BGO) с эффективной площадью 0.3 $м^2$ [7].

3 Алгоритм поиска событий

Нами разработан собственный алгоритм поиска событий в архивных данных эксперимента SPI/INTEGRAL, полученных за период с 12 июля 2003 года по 23 января 2010 года.

Гамма-спектрометр SPI/INTEGRAL позволяет регистрировать отдельные гамма-кванты, записывая момент регистрации фотона, его энергию и номер детектора, в котором произошла регистрация. Поиск событий проводился в энергетическом диапазоне [20–650, 2000–8000] кэВ. Диапазон [650–2000] кэВ был исключен, поскольку значительная часть отсчетов в этом диапазоне связана с шумом электроники. Отбор событий производился на масштабах времени 0.001, 0.01, 0.1, 1 и 10 сек с порогами значимости 20, 6, 5, 5 и 4 σ, соответственно. Порог в 20 σ для интервала 0.001 сек выбран таким образом, чтобы за исследуемый период времени минимизировать количество флуктуаций так, чтобы при начальном отборе получить только события большой интенсивности. Только такие короткие события длительностью 0.001 сек можно было бы обнаружить в данных других экспериментов (с более грубым временным разрешением).

При поиске событий использовалась сумма трех типов отсчетов. SGL – «обычный» отсчет, зарегистрированный в одном детекторе. PSD – отсчет, зарегистрированный в одном детекторе, форма импульса которого подтверждает его фотонную природу. DBL – отсчеты, которые зарегистрированы одновременно в двух различных детекторах вследствие комптоновского рассеяния исходного фотона внутри одного из детекторов.

Всего данным алгоритмом отобрано более ста тысяч событий на различных масштабах времени (от 1 мс до 10 сек). Для каждого обнаруженного события были построены кривая блеска и энергетическая диаграмма, проанализировано распределение отсчетов по детекторам, что затем было использовано для классификации событий.

4 Классификация обнаруженных событий

Для классификации событий, обнаруженных в эксперименте SPI, были использованы данные антисовпадательной защиты SPI-ACS и телескопа IBIS/ISGRI, также размещенных на обсерватории INTEGRAL.

Выделено три класса событий: флуктуации, кандидаты в «реальные» гамма-события (например, гамма-всплески) и 3 типа инструментальных явлений, связанных с взаимодействием детектора с заряженными частицами (взаимодействия с пучками электронов, протонами и галактическими космическими лучами высоких энергий). События типа «флуктуации», как правило, были обнаружены на пороге значимости и отсутствуют в данных других космических телескопов (в первую очередь – в данных экспериментов IBIS/ISGRI и SPI-ACS) и поэтому исключались из анализа.

Классификация построена на основе следующих критериев:

Длительность. Для космических гамма-всплесков значение этого параметра обычно лежит в пределах (0.1, 100) сек, для гамма-всплесков земного происхождения – в пределах (0.1, 1) мс, для вспышек источников SGR и AXP – в интервале (0.01, 3) сек. События, связанные с взаимодействием детекторов с заряженными частицами, в зависимости от типа имеют длительность от долей миллисекунд до долей секунды.

Вид энергетического спектра. В качестве параметра, характеризующего жесткость энергетического спектра, использовалось отношение отсчетов в диапазоне (100, 1000) кэВ к отсчетам в диапазоне (20, 100) кэВ. Для вспышек источников SGR и AXP значение данного параметра значительно меньше единицы. Спектры гамма-всплесков отличаются большим разнообразием – значение параметра жесткости может изменяться в широких пределах и составляет, в среднем, около единицы. Для

взаимодействий детекторов с заряженными частицами, в зависимости от типа, значение этого параметра либо значительно меньше единицы, либо значительно больше единицы.

Распределение отсчетов по детекторам. Для количественной оценки распределения отсчетов события по детекторам использовалось отношение максимальной скорости счета в одном детекторе к среднему значению скорости счета. Для «реальных» гамма-событий (гамма-всплесков, вспышек источников SGR и AXP), находящихся в поле зрения телескопа SPI (в том числе, для событий на краю поля зрения), значение этого параметра, как правило, лежит в интервале (1–3). Для событий второго типа взаимодействий детекторов с заряженными частицами значение этого критерия значительно превышает значение 3.

Характер спектральной эволюции служил в качестве дополнительного критерия для отбора событий, связанных с взаимодействием детекторов с заряженными частицами. Для большинства «реальных» событий характерна эволюция энергетического спектра – от жесткого к мягкому. Эволюция спектра событий, связанных с взаимодействием детекторов SPI с (предположительно) протонами, имеет, как правило, прямо противоположный характер. События, связанные с взаимодействием детекторов SPI с (предположительно) галактическими космическими лучами высоких энергий, наблюдаются в виде спектральных линий с энергиями 55, 64, и 198 кэВ, соответствующих ядерным реакциям захвата тепловых нейтронов ядрами Ge [8], когда нейтроны рождаются в результате каскадных реакций. Для отбора событий такого типа также введен параметр, представляющий собой отношение количества отсчетов в интервале (0, 50) мс относительно триггера к отсчетам в интервале (–50, 0) мс в узком диапазоне энергий (195, 201) кэВ. Данный критерий применялся лишь для отбора событий на масштабах времени 10 и 100 мс, поскольку события такого типа имеют длительность, в среднем, около 50 мс.

Особенности темпа регистрации. Большая часть вспышек источников SGR и AXP наблюдается в течение активности соответствующего источника, которая длится около месяца. Крайне неравномерный темп регистрации характерен для событий, связанных с взаимодействием детекторов SPI с протонами и пучками электронов. Темп регистрации гамма-всплесков – равномерный. Это связано с их космологической природой.

Обнаружение события в данных SPI-ACS, IBIS/ISGRI и других экспериментов. Большая часть (более 90%) «реальных» гамма-событий (подтвержденных другими космическими экспериментами) также обнаружена в данных IBIS/ISGRI, и около трети событий – в данных SPI-ACS. Обнаружение и локализация события в данных IBIS/ISGRI являются надежным признаком «реального» гамма-события. Часть событий, связанных с взаимодействием детекторов SPI с пучками электронов (около 30% событий) и галактическими космическими лучами высоких энергий (около 70% событий), обнаружены в данных SPI-ACS. Для событий-кандидатов в реальные гамма-события проведен поиск подтверждений в известных каталогах гамма-транзиентов [9–11].

5 Результаты

Рассмотрим детально свойства различных типов обнаруженных событий.

5.1 Флуктуации

Для этого типа характерно квазиравномерное распределение отсчетов по детекторам и по энергиям. Профиль кривой блеска также не выявляет каких-либо особенностей. В эту категорию входят также срабатывания, которые не удается отнести к другому классу событий. Все события этого типа исключены из дальнейшего анализа.

5.2 Взаимодействие детекторов SPI с высокоэнергичными протонами

Срабатывания данного типа легко идентифицировать, поскольку они обладают сразу несколькими наблюдательными особенностями. Пример такого события представлен на Рис. 1.

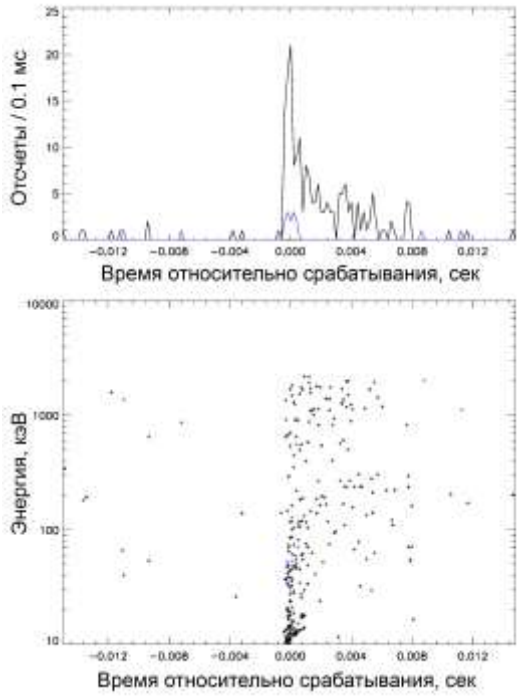


Рисунок 1 Событие, связанное с взаимодействием детекторов SPI с высокоэнергичным протоном. В верхней части рисунка – кривая блеска в диапазоне 20 кэВ – 8 МэВ. В нижней части рисунка – соответствующая кривой блеска энергетическая диаграмма. Черным цветом показаны SGL- и PSD-отсчеты, синим

DBL-отсчеты

Длительность составляет в среднем около 10 мс. Форма кривой блеска обычно несимметричная и характеризуется быстрым ростом и медленным экспоненциальным спадом. Главная особенность – крайне неравномерное распределение отсчетов по детекторам. Энергетический спектр, как правило, жесткий, с резким обрывом на 2 МэВ, причем жесткость спектра растет со временем. Темп регистрации событий этого типа неравномерный – наблюдается несколько достаточно коротких эпизодов активной регистрации длительностью 1–2 дня каждый, в промежутках между которыми события не регистрируются.

Мы предполагаем, что эти события связаны с взаимодействием детекторов с высокоэнергичными частицами космических лучей (вероятно, протонами), которые генерируют ливень вторичных частиц в одном из германиевых детекторов SPI, которые затем им же и регистрируются.

5.3 Взаимодействие детекторов SPI с пучками электронов

Пример события данного типа приведен на Рис. 2. Длительность в большинстве случаев не превышает 1–2 мс, однако встречаются более длинные события, состоящие из отдельных миллисекундных импульсов. Форма импульсов в большинстве случаев симметричная. Энергетический спектр – мягкий, с завалом на 100 кэВ. Распределение отсчетов по детекторам – равномерное. Около трети событий обнаружено в данных SPI-ACS и представляют собой короткие импульсы длительностью 50 мс.

Темп регистрации событий неравномерный: наблюдается несколько периодов активной регистрации длительностью около месяца каждый. В промежутки времени между этими эпизодами события практически не регистрируются. Темп регистрации периодичен и меняется с периодом 3 сут (период обращения обсерватории INTEGRAL вокруг Земли). Вероятно, что эти события регистрируются в те моменты, когда орбита обсерватории пересекает хвост магнитосферы Земли, и связаны с взаимодействием детекторов с пучками электронов внешнего радиационного пояса.

5.4 Спектральные линии 53 кэВ, 66 кэВ, 198 кэВ

Данный тип событий связан с кратковременным значительным увеличением скорости счета в фоновых спектральных линиях 198 кэВ, 53 кэВ и 66 кэВ, вызванных следующими ядерными реакциями захвата тепловых нейтронов:

$^{70}Ge+n>^{71m}Ge$ (время жизни ≈ 20.4 мс)$>^{71}Ge+\gamma$ двухступенчатое излучение 175+23=198 кэВ [8];

$^{72}Ge+n>^{73m}Ge$ (время жизни 0.5 с)$>^{72}Ge+\gamma$ 53 кэВ [8].

$^{72}Ge+n>^{73m}Ge>^{72}Ge+\gamma$ двухступенчатое излучение 53+13=66 кэВ [8].

Длительность транзиентного излучения в линии 198 кэВ составляет десятки мс. Транзиентное излучение в спектральных линиях 53 кэВ и 66 кэВ наблюдается не во всех случаях. Длительность интенсивного излучения в этих линиях может составлять несколько секунд.

Событиям этого класса в 25% случаев сопутствует «насыщение» германиевых детекторов SPI – отсутствие сигнала в одном или нескольких соседних детекторах в течение нескольких секунд. Природа насыщения не ясна.

Темп регистрации – квазиравномерный на уровне около 2 событий в сутки.

Более 70% событий наблюдаются также в данных SPI-ACS. В [12] показано, что кристаллы BGO, из которых состоит SPI-ACS, в результате взаимодействия с космическими лучами (ядерные реакции скалывания) испускают вторичные нейтроны, которые, в свою очередь, термализуются и захватываются ядрами Ge детекторов эксперимента SPI. Вероятно, рассмотренные события связаны с взаимодействием наиболее энергичных галактических космических лучей, генерирующих мощный каскад вторичных частиц в SPI-ACS и регистрируемых затем детекторами SPI.

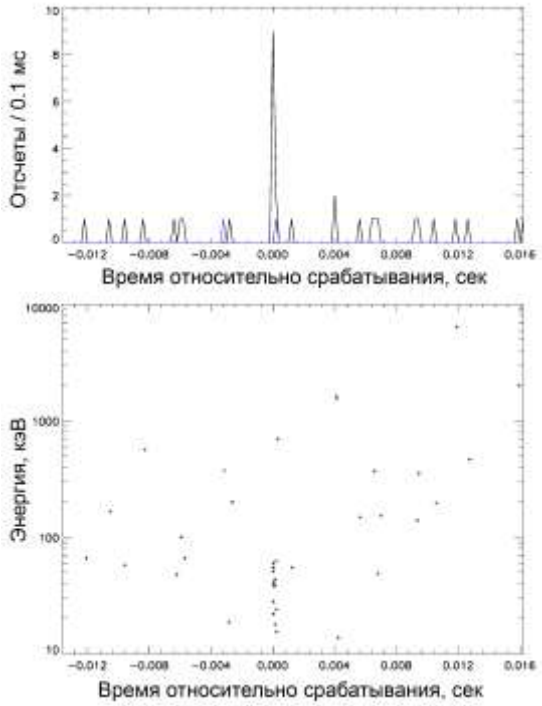


Рисунок 2 Событие, связанное с взаимодействием детекторов SPI с пучком электронов магнитосферы Земли. То же, что на Рис. 1

5.5 Кандидаты в гамма-всплески земного происхождения

Для исследования диффузного рентгеновского фона обсерваторией INTEGRAL был произведен ряд наблюдений, когда в апертуру телескопов на

борту обсерватории попадала Земля. Мы использовали эти данные общей длительностью 496 кс для поиска гамма-всплесков земного происхождения (TGF). Всего было отобрано 28 кандидатов на основе известных свойств TGF: длительность ≤1 мс, энергетический спектр – жесткий, с регистрацией фотонов свыше 1 МэВ, форма кривой блеска – симметричная, спектральная эволюция отсутствует, распределение зарегистрированных отсчетов по детекторам SPI – квазиравномерное. Один из кандидатов представлен на Рис. 3. Детальное исследование кандидатов в TGF, обнаруженных в данных SPI/INTEGRAL, см. в [13].

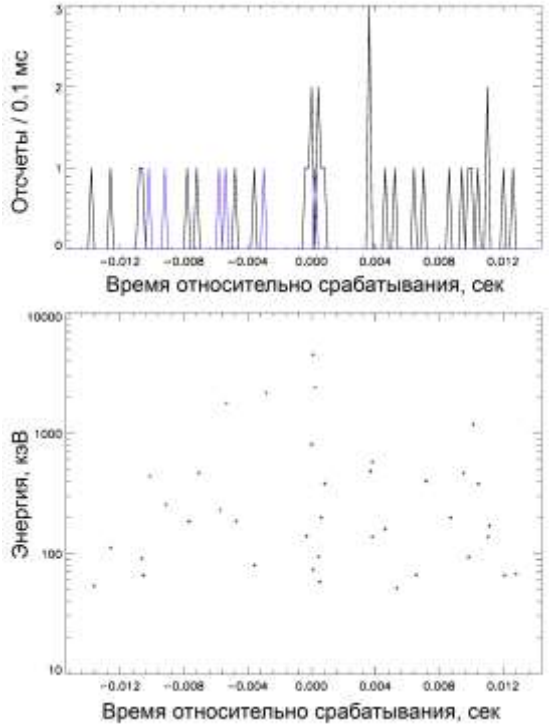


Рисунок 3 Пример кандидата в гамма-всплеск земного происхождения (TGF). То же, что на Рис.1

5.6 Вспышки источников SGR 1806-20 и AXP 1E_1547.0-5408

Были обнаружены 223 вспышки источника мягкого повторного гамма-излучения SGR 1806-20 (Рис. 4) и 23 вспышки аномального рентгеновского пульсара AXP 1E_1547.0-5408, которые были обнаружены и локализованы в эксперименте IBIS/ISGRI (см. таблицы 1–3 в [1]). Часть событий также обнаружена в данных SPI-ACS.

Кроме того, составлен список из 90 кандидатов во всплески источников повторного мягкого излучения (SGR и AXP), отобранных в соответствии с наблюдаемыми свойствами подтвержденных всплесков источников типа SGR, а именно: длительность отдельных импульсов события находится в пределах (0.01–3) сек; доля фотонов с энергией выше 200 кэВ пренебрежимо мала (мягкий спектр); распределение по детекторам близко к равномерному.

5.7 Космические гамма-всплески

Было обнаружено 48 космических гамма-всплесков, подтвержденных другими космическими экспериментами (см. таблицы 4–6 в [1]). На Рис. 5 представлены кривая блеска и энергетическая диаграмма для гамма-всплеска GRB 050525.

Помимо подтвержденных гамма-всплесков было отобрано 160 кандидатов (см. таблицу 7 в [1]), из них 151 событие относится к коротким всплескам с длительностью менее 2 сек. Кандидаты отбирались в соответствии с наблюдаемыми свойствами гамма-всплесков, подтвержденных другими космическими телескопами, а именно: длительность отдельных импульсов события – более 5 мс; жесткий энергетический спектр у событий с длительностью менее 2 сек (данный критерий также позволяет отсеивать события, связанные с активностью источников повторного мягкого гамма-излучения SGR); распределение по детекторам близко к равномерному.

Детали исследования космических гамма-всплесков, зарегистрированных в эксперименте SPI/INTEGRAL, в том числе спектральный анализ, исследование спектральной эволюции и распределения гамма-всплесков по длительности, см. в работах [1, 14].

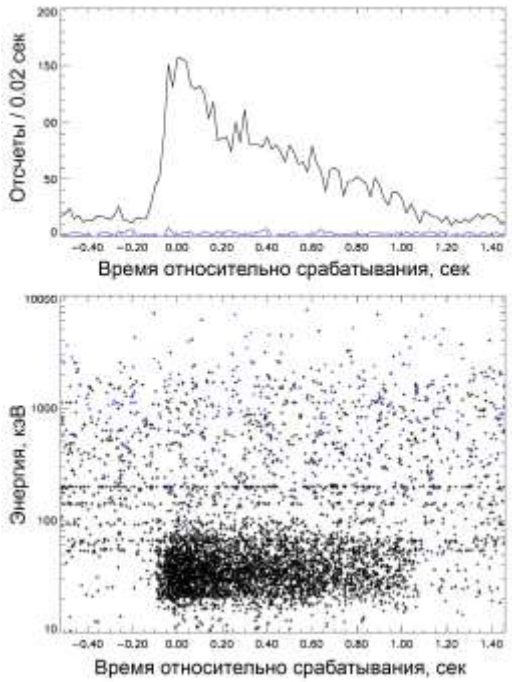


Рисунок 4 Пример вспышки источника SGR 1806-20. То же, что на Рис. 1

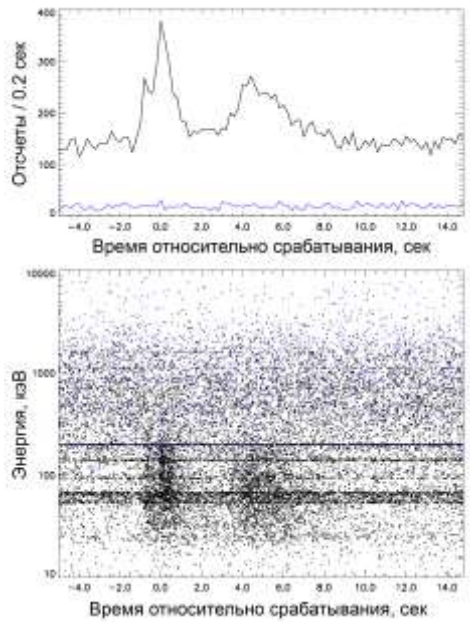


Рисунок 5 Космический гамма-всплеск GRB 050525. То же, что на Рис. 1

6 Заключение

Проведено комплексное исследование архивных наблюдательных данных гамма-спектрометра SPI, накопленных за 7 лет работы обсерватории INTEGRAL.

Предложен алгоритм поиска транзиентных событий на различных временных масштабах от 1 мс до 10 сек, с помощью которого было обнаружено более 100 000 событий. Разработана методика классификации обнаруженных событий на основе ряда критериев, которая может применяться автоматически непосредственно после обнаружения события в наблюдательных данных.

Выделены три класса событий: флуктуации; «реальные» гамма-события (гамма-всплески космического (GRB) и земного (TGF) происхождения, вспышки источников SGR и AXP); три типа инструментальных явлений, связанных с взаимодействием детектора с заряженными частицами (взаимодействия с пучками электронов, протонами, и галактическими космическими лучами высоких энергий).

Составлены каталоги космических гамма-всплесков и вспышек источников SGR 1806-20 и AXP 1E_1547.0-5408.

Благодарности

Работа поддержана грантом РФФИ (проект 16-32-00489 мол_а) и частично грантом РФФИ 17-02-01388.

Литература

[1] Minaev, P.Yu., Pozanenko, A.S., Molkov, S.V., Grebenev, S.A.: Catalog of Short Gamma-ray Transients Detected in the SPI/INTEGRAL Experiment. Astronomy Letters, 40, p. 235 (2014)

[2] Gurevich, A.V., Milikh, G.M., Roussel-Dupre, R.: Runaway Electron Mechanism of Air Breakdown and Preconditioning During a Thunderstorm. Physics Letters A, 165, pp. 463-468 (1992)

[3] Briggs, M.S., Xiong, S., Connaughton, V. et al.: Terrestrial Gamma-ray Flashes in the Fermi Era: Improved Observations and Analysis Methods. J. of Geophysical Research (2013). doi: 10.1002/jgra.50205

[4] Vedrenne, G., Roques, J.-P., Schönfelder, V. et al.: SPI: The Spectrometer Aboard INTEGRAL. Astronomy and Astrophysics, 411, L63–L70 (2003)

[5] Winkler, C., Courvoisier, T. J.-L., Di Cocco, G. et al.: The INTEGRAL Mission. Astronomy and Astrophysics, 411, L1–L6 (2003)

[6] Vedrenne, G., Roques, J.-P., Schönfelder, V. et al.: SPI: The Spectrometer Aboard INTEGRAL. Astronomy and Astrophysics, 411, L63–L70 (2003)

[7] von Kienlin, A., Beckmann, V., Rau, A. et al.: INTEGRAL Spectrometer SPI's GRB Detection Capabilities. GRBs Detected Inside SPI's FoV and with the Anticoincidence System ACS. Astronomy and Astrophysics, 411, L299–L305 (2003)

[8] Weidenspointner, G., Harris, M.J., Sturner, S. et al.: MGGPOD: a Monte Carlo Suite for Modeling Instrumental Line and Continuum Backgrounds in Gamma-Ray Astronomy. Astrophys. J. Suppl., 156 (69), astro-ph/0408399 (2005)

[9] Chelovekov, I.V., Grebenev, S.A.: Hard X-ray Bursts Recorded by the IBIS Telescope of the INTEGRAL Observatory in 2003-2009. Astronomy Letters. 37, p. 597, arXiv:astro-ph.HE/1108.2421 (2011)

[10] IBAS IBIS/ISGRI triggers: http://ibas.iasf-milano. inaf.it/

[11] Hurley, K.: Masterlist, http://www.ssl.berkeley. edu/ipn3/chronological.txt

[12] Jean, P., von Ballmoos, P., Vedrenne, G., Naya, J.E.: Performance of Advanced Ge-spectrometer for Nuclear Astrophysics. Gamma-Ray and Cosmic-Ray Detectors, Techniques, and Missions. Ed. by B.D. Ramsey, T.A. Parnell. Vol. 2806 of Society of Photo-Optical Instrumentation Engineers (SPIE) Conference Series, p. 457 (1996)

[13] Минаев, П.Ю., Позаненко, А.С., Гребенев, С.А., Мольков, С.В.: Возможности и оценки наблюдений гамма-всплесков земного происхождения (TGF) космической обсерваторией INTEGRAL. (в печати)

[14] Minaev, P.Yu., Grebenev, S.A., Pozanenko, A.S. et al.: GRB 070912 – A gamma-ray Burst Recorded from the Direction to the Galactic Center. Astronomy Letters, 38, pp. 613 (2012)

Разработка каталога идентификации двойных звёзд ILB

© Н.А. Скворцов © Л.А. Калиниченко

ФИЦ «Информатика и управление» РАН
Москва, Россия

© А.В. Карчевский © Д.А. Ковалева © О.Ю. Малков

Институт астрономии РАН
Москва, Россия

nskv@mail.ru leonidandk@gmail.com

geisterkirche@gmail.com dana@inasan.ru malkov@inasan.ru

Аннотация. Двойные и кратные системы звёзд наблюдают с использованием разных методов и инструментов. Каталоги двойных звёзд определённых наблюдательных типов независимы друг от друга и используют свои системы идентификации звёзд. Также компоненты двойных соотнесены с идентификаторами обзоров и каталогов одиночных звёзд. Задача перекрёстной идентификации двойных звёзд различных наблюдательных типов, а также обзоров неба нетривиальна и связана с разрешением разного рода конфликтов. Она требует не просто объединения списков существующих идентификаторов для конкретных звёзд, а отождествления компонентов кратных систем по астрометрическим и астрофизическим параметрам для дальнейшего соотнесения идентификаторов определённым компонентам и друг другу. В данной статье описана разработка средств для создания каталога идентификаторов двойных звёзд ILB, включающая процедуру перекрёстного отождествления систем, их компонентов и пар всех наблюдательных типов. Работа является продолжением исследований методов отождествления двойных и кратных систем.

Ключевые слова: система идентификации, двойные звёзды, разрешение сущностей.

Development of Identification List of Binaries ILB

© N.A. Skvortsov © L.A. Kalinichenko

FRC «Computer Science and Control» RAS,
Moscow, Russia

© A.V. Karchevsky D.A. Kovaleva © O.Yu. Malkov

Institute of Astronomy RAS,
Moscow, Russia

nskv@mail.ru leonidandk@gmail.com

geisterkirche@gmail.com dana@inasan.ru malkov@inasan.ru

Abstract. Binary and multiple stellar systems have been observed using various methods and tools. Catalogs of binaries of different observational types are independent and use inherent star identification systems. Moreover, components of stellar systems refer identifiers of surveys and catalogs of single stars. The problem of cross-identification of binary stars of different observational types as well as sky surveys is non-trivial and related to resolution of various kinds of conflicts. It requires not only combining lists of existing identifiers of specific stars, but matching components of multiple systems according to astrometric and astrophysical parameters for further referring of identifiers to matched components and to each other. This paper describes development of tools for creating the Identification List of Binaries (ILB) including cross-matching of systems, their components and pairs of all observational types. This work continues research of binary and multiple systems matching methods.

Keywords: identification systems, binary stars, entity resolution.

1 Введение

Двойные звезды довольно многочисленны и составляют значительную часть звездной популяции галактики (от 20% до 90%, по разным оценкам для разных выборок). Значительная часть двойных звезд на самом деле являются системами большей кратности. Самый большой каталог визуальных двойных звезд WDS [1] содержит более 100 000 пар, из которых 25000 в системах с кратностью три и больше.

Есть веские основания считать, что компоненты двойной звезды формируются одновременно и в дальнейшем эволюционируют параллельно, оставаясь в системе. Фактором, определяющим ход эволюции, является распределение первоначальной массы между компонентами. Поэтому для определения принадлежности компонентов системе необходимо оценивать общность их эволюции.

Двойные звезды подразделяют на несколько типов в зависимости от способа их наблюдения. Для каждого типа наблюдений формировались отдельные каталоги с собственными наборами наблюдаемых параметров. Основные наблюдательные типы составляют визуальные, астрометрические, орбитальные, интерферометрические, затменно-переменные, спектральные двойные.

Среди визуальных пар различают оптические и физические двойные. И те, и другие пары можно найти в каталогах визуальных двойных, в частности, в WDS. Оптические пары состоят из весьма далеких и несвязанных в пространстве звезд, проецирующихся на небесную сферу близко друг к другу в направлении наблюдения. Физические пары представляют собой близко расположенные в пространстве компоненты, связанные силами тяготения, обращающиеся вокруг общего центра масс по законам Кеплера. Если наблюдения продолжаются достаточно долго, может прослеживаться полное обращение звезды. В результате их наблюдений определяют взаимное угловое расстояние компонентов и позиционный угол. Это самая многочисленная группа известных двойных звёзд. Основными каталогами визуальных двойных являются WDS, CCDM [2], Tycho [3].

Если один из двух компонентов не виден по тем или иным причинам, двойственность можно обнаружить по изменению положения на небе второго компонента. В таком случае говорят об астрометрических двойных звёздах. Основными каталогами таких звёзд являются два каталога Makarov and Kaplan. Затменно-переменные двойные звезды представляют собой пары, радиус обращения которых сравним с размерами самих звезд, а плоскости орбит этих звезд и луч зрения наблюдателя практически совмещаются. Эти звезды обнаруживаются явлениями затмений, проявляющимися периодическим падением яркости наблюдаемой звезды. В результате наблюдений определяются параметры кривых блеска, отражающие закономерности изменения яркости звезды со временем. Основными каталогами затменных являются ОКПЗ [4] и CEV2. Интерферометрические двойные звезды наблюдаются при помощи Фурье-анализа изображений телескопов, увеличивающего разрешающую способность до дифракционного предела. Обнаруженные таким образом двойные представлены в каталоге INT4. Спектральные двойные звезды представляют собой пары, обращающиеся в плоскости, слабо наклоненной к направлению луча зрения наблюдателя. Они обнаруживаются при спектроскопических наблюдениях лучевых скоростей. Линии в спектрах таких звезд регулярно смещаются или раздваиваются из-за эффекта Доплера, что свидетельствует о двойственности звезды. В результате наблюдений определяют кривые лучевой скорости, амплитуду и период колебаний. Основным источником данных о спектральных двойных является каталог SB9 [5]. Существуют и другие наблюдательные типы двойных и специализированные каталоги.

В разных сообществах исследователи специализировались на различных способах наблюдения двойных, поэтому принципы построения каталогов разных наблюдательных типов двойных никоим образом не согласовывались. В дополнение к существенной неоднородности, рождающей конфликты, проявляющиеся при интеграции каталогов, в большинстве каталогов созданы собственные системы идентификации двойных. Некоторые из каталогов также содержат ссылки на идентификаторы соответствующих наблюдений в обзорах одиночных звёзд.

База данных двойных (BDB) [6], созданная авторами данной статьи, включает данные о двойных и кратных системах всех наблюдательных типов, собранные из разных каталогов, с некоторыми общими параметрами и отсылкой на оригинальные каталоги по идентификаторам. При создании базы была осознана необходимость разработки специализированной системы идентификации (BSDB) [7], учитывающей идентификацию компонентов, пар внутри систем и кратных систем в целом. Однако эта система сама по себе не решает изначальную неоднородность идентификаторов и требует аккуратного отождествления с идентификаторами разных систем идентификации.

С этой целью отдельно создаётся каталог идентификации двойных ILB, объединяющий в себе перекрёстные значения целого ряда систем идентификации. Для создания этого каталога недостаточно просто свести в таблицу необходимые идентификаторы, исходя из совпадения уже присутствующих в оригинальных каталогах перекрёстных идентификаций, так как эта работа сталкивается с множеством конфликтных ситуаций. Разрешение возникающих конфликтов основано на астрометрическом и астрофизическом подходах к отождествлению компонентов, пар и систем звёзд.

Разработан алгоритм отождествления многокомпонентных сущностей, позволяющий корректно соотносить данные различных наблюдений кратных звёздных систем. Результат его работы используется для разрешения

конфликтов между идентификаторами разных систем идентификации.

Основной целью данной работы является описание реализации каталога идентификаторов двойных звезд ILB (Identification List of Binaries), который объединяет идентификаторы двойных и кратных звезд всех наблюдательных типов. Формирование каталога потребовало создания инструментария, предназначенного для заполнения и поддержки базы идентификаторов в актуальном состоянии, а также расширения её по мере необходимости. Некоторые каталоги двойных и обзоры являются периодически обновляемыми, соответственно база идентификаторов должна обновляться вместе с ними.

Основные принципы алгоритма, лежащего в основе инструментов создания ILB, приведены в разделах 2 и 3. Раздел 4 посвящён вопросам реализации программных инструментов и результатам создания каталога.

2 Каркас для отождествления двойных и кратных звёзд

Общий подход к отождествлению одиночных и многокомпонентных сущностей включает построение множеств кандидатов на идентификацию для каждой сущности или её компонента и применение набора критериев отождествления, ограничивающих такие множества кандидатов. Критерии отождествления формируются на основании знаний предметной области, ограничивающих интерпретацию объектов, и определяются унифицированным образом, принимая в качестве аргументов отождествляемые сущности, а не отдельные параметры.

Первыми применяются критерии, которые обычно наиболее сильно ограничивают начальное множество кандидатов, для этой цели критериям может присваиваться приоритет. В остальном, они применяются в произвольном порядке тогда, когда присутствуют все требуемые данные о сущностях для их применения. В случае несоответствия применяемому критерию отождествляемая сущность исключается из списка кандидатов на идентификацию с текущей сущностью.

Если данные о сущности, необходимые для применения критерия, отсутствуют, по нему не может быть ограничено множество кандидатов. Таким образом, присутствие данных о специфических атрибутах даже для небольшой части сущностях предметной области позволяет применять связанные с этими атрибутами дополнительные критерии отождествления, а отсутствие определённых данных влияет только на применимость связанных с ними критериев, но не на возможность отождествления сущностей в целом.

Подходы, используемые для разрешения сущностей, включают в себя различные критерии сходства для подмножеств значений атрибутов и структур графов, которые позволяют оценивать тождественность многокомпонентных сущностей. Критерии, основанные на знаниях о сущностях предметной области, могут ограничивать возможные значения атрибутов или их сочетание у одной сущности или вводить ограничения на изменчивость значений атрибутов отождествляемых объектов. Атрибуты могут иметь константные значения для определённой сущности, либо менять значения в рамках определённых ограничений, выход за которые будет означать, что объекты описывают разные сущности.

Графовые критерии могут включать правила отождествления, основанные на ограничениях структуры многокомпонентных сущностей, или делать выводы об отождествлении многокомпонентных сущностей или компонентов на основе уже установленных идентификаций других компонентов.

2.1 Унификация данных в предметной области

Типы сущностей задаются концептуальной схемой предметной области. Она определяет абстрактные типы данных, описывающие структурированное представление информации о сущностях, их ограничениях, а также спецификации поведения сущностей.

Для каждого типа сущности X в концептуальной схеме создается абстрактный тип данных $T_X[a_1, \dots, a_i, \dots]$, содержащий атрибуты a_i для описания характеристик, которые сущность может иметь в данной предметной области. В предметной области двойных и кратных систем звёзд все астрономические объекты и связь их друг с другом рассматриваются в терминах трёх основных типов сущностей: кратных систем звёзд в целом, их отдельных компонентов и пар компонентов. Особым типом является также тип идентификатора, который связывается с одним из типов объектов: компонентом, парой или системой.

Набор источников данных D_j (обзоров и каталогов) хранит данные о сущностях определенного типа, структурированных как кортежи $X_j[a_{1j}, \dots, a_{ij}, \dots]$, содержащие атрибуты a_{ij}, относящиеся к характеристикам сущности X в типе T_X. В разных астрономических каталогах представления X_j могут быть разными. В частности, записи каталога WDS описывают данные о парах компонентов систем, в то время как каталог CCDM рассматривает в качестве записей данные о визуальных компонентах систем. Поэтому сбор данных из нескольких источников производится с одновременным преобразованием данных в унифицированное представление в терминах концептуальной схемы предметной области, в котором и происходит дальнейший

анализ, в частности перекрестное отождествление сущностей. Преобразование данных требует построения отображения M исходных данных X_j в соответствующее значение типа T_X:

$$M_{jX}: X_j \to T_X$$

Функция отображения представляет собой набор правил преобразования подмножеств атрибутов из исходного представления в источниках данных в атрибуты типа концептуальной схемы. Для унификации представления данных и повышения его качества отображение также включает функции стандартизации в качестве правил очистки и унификации, применяемые к атрибутам типов X_j. Например, параметр прямого восхождения в каталоге WDS представлен в формате HHMMSS.ss, в каталоге SB9 – HHMMSSSSS, а в TDSC – и вовсе в градусной мере. Поэтому реализуется преобразование их к общему виду в концептуальной схеме. Записи каталога INT4 в целом нетривиально построены, и в качестве функций отображения требуют непростых преобразований, формирующих унифицированное представление данных в терминах концептуальной схемы.

С применением функций отображения данные из неоднородных источников преобразуются к унифицированному представлению, соответствующему концептуальной схеме предметной области, и в дальнейшем обрабатываются только в нём. Соответственно алгоритмы анализа данных разрабатываются над концептуальной схемой предметной области.

2.2 Организация работы с критериями отождествления сущностей

После унификации представления данных в концептуальной схеме начинается процесс отождествления сущностей. Для каждого объекта формируется множество кандидатов на идентификацию $C: \{T_X\}$. К каждому кандидату прикрепляется множество флагов (F), используемых в ходе работы алгоритмов.

Множество кандидатов на идентификацию для конкретного объекта строится с использованием набора критериев, построенных на ограничениях предметной области, относящихся к сущности X. Применение определённого критерия отождествления производится с помощью функции:

$$T_X \times \{T_X\} \to \{T_X\}$$

Рассматривается объект типа T_X, одиночный или многокомпонентный, представляемый в первом аргументе функции. Также доступно множество возможных кандидатов на идентификацию $C: \{T_X\}$ в качестве второго аргумента. Функция отождествляет каждый объект из множества кандидатов с рассматриваемым объектом по данному определённому критерию. В результате функция возвращает уменьшенный набор кандидатов, которые отвечают критерию.

Критерий k, ограничивающий множество кандидатов C на идентификацию с объектом X, определяется предикатом R_{kX}, содержащим определённое ограничение предметной области над рассматриваемым (x) и отождествляемым (c) объектами:

$$\{c \in C | R_{kX}(x, c)\}$$

Предикат может использовать значения атрибутов и флаги самих объектов, а также атрибуты и флаги их компонентов в случае, если объекты многокомпонентные. Так, для отождествления пары звёзд могут сравниваться как значения атрибутов самих пар, так и атрибутов каждого из компонентов пары. Например, в тождественных парах должны быть близкие значения собственных движений компонентов. Функция, определяющая данный критерий, сравнивает значения атрибутов собственного движения рассматриваемой пары и пары-кандидата, либо при отсутствии значений у пар может получать данные о собственном движении из атрибутов компонентов пар и сравнивать их.

Уникальная идентификация – особая ситуация, когда множество C после применения всех возможных критериев содержит единственный объект из определённого источника данных в качестве кандидата на идентификацию. В большинстве случаев это означает, что наиболее вероятный кандидат найден. С таким объектом связывается специальный флаг. При этом возникают условия для применения критериев, использующих этот флаг. Они применяются к элементам графа, связанного с данным объектом, для идентификации других компонентов в многокомпонентном объекте.

3 Критерии отождествления кратных систем

Процесс отождествления многокомпонентных сущностей разделяется на несколько взаимодействующих этапов по типам сущностей, а также по вложенности структуры. При отождествлении двойных и кратных звёздных систем процесс начинается с отождествления отдельных компонентов систем как одиночных сущностей. При этом используются данные каталогов визуальных двойных звёзд и обзоры одиночных звёзд, ссылки на идентификаторы которых приводятся в каталогах двойных. Затем на основании результатов первого этапа отождествляются визуальные пары звёзд. На следующем этапе рассмотрение дополняется более тесными парами других наблюдательных типов. Отождествление систем в целом является следствием отождествления всех их компонентов и пар. Наконец, существующие идентификаторы систем, пар и компонентов сопоставляются с использованием результатов предыдущих этапов.

В [8] был кратко описан состав знаний, используемых для формирования критериев отождествления, используемых на каждом этапе.

Для составления критериев отождествления компонентов систем как одиночных объектов рассмотрены следующие ограничения предметной области, связанные одиночными звёздными объектами:

- близость отождествляемых компонентов по координатам;
- учёт изменения координат в разные наблюдательные эпохи по причине прецессии земной оси и собственного движения компонентов;
- близость направления и скорости собственного движения звёзд;
- близость тригонометрического параллакса компонентов, говорящего об их расстоянии от наблюдателя;
- близость значений блеска или разницы показателей цвета при условии известных фотометрических полос пропускания и с учетом чувствительности к низким и высоким пределам величины;
- учет возможной переменности звёзд при сравнении блесков;
- сходство эволюционных статусов звёзд;
- сходство спектральной классификации звёзд.

На основе выборки кандидатов на идентификацию компонентов систем составляется граф, содержащий пары из различных каталогов двойных. Для составления критериев отождествления широких визуальных пар в полученном графе рассматривались следующие ограничения:

- если в парах выделены компоненты, кандидаты на идентификацию пар формируются на основе множеств кандидатов на идентификацию компонентов, сформированных с применением вышеописанных критериев;
- проверяется близость значений позиционного угла и углового расстояния;
- предельная разность позиционных углов с учётом углового расстояния для оценки возможного вращения компонентов в паре;
- учёт минимальной оценки периода вращения пары;
- учёт данных орбитального движения;
- близость направления и скорости собственного движения пары в целом или компонентов в паре;
- возможность существенного различия собственных движений компонентов в оптических парах;
- близость разницы блесков компонентов в паре;
- учёт эффективного углового разрешения (расстояние в паре меньше предела различения объектов в каталоге);
- сходство химического состава и эволюционного статуса компонентов в физических парах;
- учёт известного наблюдательного типа пары.

Идентификация многокомпонентных объектов зависит от идентификации их компонентов. Уникальная идентификация всех компонентов означает идентификацию всего объекта. Уникальная идентификация пар (дуг в графах) означает идентификацию обратных пар. Уникальная идентификация многокомпонентных объектов как целого означает, что идентификация всех компонентов должна быть разрешена среди кандидатов, принадлежащих идентифицированному объекту.

В действительности, наиболее сильные ограничения дают критерии отождествления компонентов на основании координат, а пар – на основании расстояния между компонентами и позиционных углов. При разработке каталога ILB для отождествления кратных систем использовался сокращённый набор критериев. В подавляющем большинстве случаев для идентификации отдельных компонентов достаточно применения критериев отождествления, связанных с координатами и взаимном положении компонентов. Остальные критерии не дают существенного эффекта при решении данной задачи, хотя могут быть использованы для проверки качества данных. Задача поиска ошибок в оригинальных каталогах на данный момент не ставилась.

Для отождествления компонентов, в первую очередь, ограничивается область координат отождествляемых объектов для наибольшего ограничения количества кандидатов на идентификацию компонентов. Отождествление пар также проводится с применением наиболее сильных критериев. Используется критерий близости позиционных углов и угловых расстояний. На данный момент упомянутые критерии реализованы совместно в виде функционала.

```
f1 = (r1 - r2) * (r1 - r2) / max(r1, r2) /
max(r1, r2) + (t1 - t2) * (t1 - t2) / max(t1,
t2)/ max(t1, t2);

f2 = (r1 - r2) * (r1 - r2) / max(r1, r2)/
max(r1, r2) + (t1 - t2 - 180) * (t1 - t2 -
180) / max(t1, t2)/ max(t1, t2);

X1 = Math.min(func1,func2)/2;
```

Здесь `r1`, `r2` — угловые расстояния между компонентами в отождествляемых парах, `t1`, `t2` — соответствующие позиционные углы. Метрика `X1` не должна превышать некоторого предела для ограничения множества кандидатов на идентификацию. Для малых расстояний между компонентами используется метрика, игнорирующая значения позиционных углов и оценивающая только координаты. Для однозначной идентификации не должно присутствовать других звёзд в окрестности,

удовлетворяющих этому критерию. Алгоритмы отождествления имеют сложность порядка $O(n^2 \log n)$.

Реализация критериев для формирования ILB сегодня ещё претерпевает изменения. В случае, если есть данные об эпохах наблюдения и собственных движениях, должен использоваться критерий сравнения координат, скорректированных по прецессии и собственному движению, что повышает качество автоматической идентификации [9]. Помимо этого, планируется проверять разность позиционных углов с учётом возможного вращения компонентов в паре [10].

Идентификаторы отождествляются по принадлежности одним и тем же компонентам, парам и системам. Зачастую один и тот же идентификатор в одном каталоге должен соответствовать компоненту (или одиночной звезде), а в другом – паре звёзд, так как угловое разрешение каталога позволяет различать эту пару визуально. Разрешение таких конфликтов происходит за счёт качественного отождествления компонентов и пар и существенно влияет на результат сопоставления идентификаторов в ILB.

Например, одним из критериев отождествления пар звёзд является требование близких значений их собственных движений.

4 Система идентификации и состав идентификаторов

Основной системой идентификации в ILB является BSDB [7], разработанная и зарегистрированная авторами статьи. Она позволяет описывать кратные системы в целом, составляющие их пары звёзд, компоненты, учитывает возможность корректировки систем при открытии новых компонентов. Она используется для обозначения кратных систем всех наблюдательных типов.

При разработке инструментов построения каталога идентификаторов двойных и кратных звёзд ILB учитывались необходимость модификации и повторного использования программ. В архитектуре инструментов учтены возможности подключения произвольных каталогов, изменения правил отождествления для разных типов сущностей. Использовался язык Java и шаблон проектирования IoC (Inversion of Control), в котором управление остаётся за каркасом, а логика работы с объектами сосредоточена в независимых и взаимозаменяемых модулях. При этом подключение каталогов или модификация алгоритмов для работы с разными наблюдательными типами астрономических объектов становятся простыми задачами. Инструмент содержит необходимый на сегодня набор методов отождествления и вспомогательные методы для работы с данными из каталогов, кэширования объектов небольшой части неба, очистки данных, сбора статистики.

ILB в настоящий момент является каталогом-основой (или мастер-каталогом) для BDB и содержит все данные BDB по кросс-идентификации более 130000 двойных и кратных систем. ILB содержит координаты и перекрёстные ссылки на следующие идентификаторы: Bayer/Flamsteed, DM (BD/CD/CPD), HD, HIP, ADS, WDS, CCDM, TDSC, GCVS, SBC9. На данный момент в списке идентификаторов ILB находится:

- Систем :136885;
- Записей о парах :313811;
- Записей о компонентах :627460.

И идентификаторов:

- HIP:36560, из которых 20701 уникальных;
- HD:29882, из которых 27917 уникальных;
- DM:121067, уникальных 105569;
- ADS:49067, уникальных 15389;
- FLAMSTEED:1622, уникальных 1529;
- BAYER:600, уникальных 548.

Заключение

Проблема сопоставления идентификаторов, используемых для обозначения двойных и кратных звёзд различных наблюдательных типов решается составлением каталога идентификации ILB.

В работе описана проблема перекрёстной идентификации кратных звёздных систем. Приводятся принципы и каркас системы разрешения многокомпонентных сущностей, и применение этих принципов для перекрёстного отождествления кратных звёздных систем, включая отождествление компонентов и пар с привлечением всех имеющихся астрометрических и астрофизических параметров объектах. Рассказано о реализации инструментария для создания и поддержки каталога ILB.

Каталог ILB содержит перекрёстную идентификацию кратных звёзд, содержащихся в каталогах двойных всех основных наблюдательных типов. База идентификаций используется обеспечивает все необходимые идентификации для базы данных двойных звёзд BDB и формирует используемые в ней идентификаторы двойных и кратных звёзд в системе идентификации BSDB.

Благодарности

Работа выполнена при поддержке РФФИ (гранты 16-07-01162, 16-07-01028).

Литература

[1] Mason, B.D., et al. The Washington Visual Double Star Catalog. VizieR on-line data catalog:

B/wds (2016). http://cdsarc.u-strasbg.fr/viz-bin/Cat?B/wds

[2] Dommanget, J., Nys, O. Catalogue of the Components of Double and Multiple Stars (CCDM). VizieR on-line data catalog: I/274 (2002). http://cdsarc.u-strasbg.fr/viz-bin/Cat?I/274

[3] Fabricius, C., Hog, E., Makarov, V.V., et al. The Tycho double star catalogue. In: Astronomy & Astrophysics, vol. 384, iss 1, pp. 180-189 (2002).

[4] Samus N. N., Durlevich O. V., Kazarovets E. V. et al. General Catalogue of Variable Stars. VizieR On-line Data Catalog: B/gcvs. (2013) http://cdsarc.u-strasbg.fr/viz-bin/Cat?B/gcvs

[5] Pourbaix, D., Tokovinin, A.A, Batten, A.H, et al. SB9: 9th Catalogue of Spectroscopic Binary Orbits. VizieR On-line Data Catalog: B/sb9 (2014) http://cdsarc.u-strasbg.fr/viz-bin/Cat?B/sb9

[6] Malkov O., Kaygorodov P., Kovaleva D. et al. Binary star database BDB: datasets and services. In: Astronomical and Astrophysical Transactions (AApTr), Vol. 28, Issue 3, pp. 235-244.(2014)

[7] D.A. Kovaleva, O.Yu. Malkov, P.V. Kaygorodov, et al. Bsdb: a new consistent designation scheme for identifying objects in binary and multiple stars Open Astronomy, vol. 24, Issue 2, pp. 185–193. (2015). doi: 10.1515/astro-2017-0218

[8] Skvortsov, N.A., Kalinichenko, L.A., Kovaleva, D.A., Malkov, O.Y. Hierarchical Multiple Stellar Systems. In: Kalinichenko L.A., Kuznetsov S.O., Manolopoulos Y. (eds) Data Analytics and Management in Data Intensive Domains. DAMDID/RCDL 2016. Communications in Computer and Information Science, vol. 706. pp 119-129. Springer, Cham (2017). doi: 10.1007/978-3-319-57135-5_9

[9] Жаров, В.Е. Сферическая астрономия. Москва (2002). http://www.astronet.ru/db/msg/1190817/node49.html

Isaeva, A.A., Kovaleva, D.A., Malkov, O.Y. Visual binaries: cross-matching and compiling of a comprehensive list. Open Astronomy, vol. 24, pp. 157–165. (2015)

Техники Семантического Веба в ОИИД

Semantic Web techniques in DID

Semantic Educational Web Portal

© Victor Telnov

National Research Nuclear University MEPhI,
Obninsk, Russia

telnov@bk.ru

Abstract. The paper deals with the pilot project devoted to the application of the knowledge graphs in the educational activities of the universities. The ontology of the curriculum and the training courses, as well as the means of authoring, enrichment and adaptation of the learning objects are considered. The visual navigation on the knowledge graphs is carried out by using the special searching widgets and smart RDF browser. Working with semantic repository and text analytics is performed on the cloud platforms via SPARQL queries and RESTful services. The software architecture in UML-notation are presented, examples of real use of the educational portal are given.

Keywords: semantic web, educational portal, ontology, knowledge graph, triplestore, RDF storage, graph database, cloud computing.

1 Introduction

Students and professors spend a lot of time and efforts finding useful information, instead of having to comprehend and interpret the learning content. It was rightly observed that the traditional web technologies (sometimes referred to as WEB 2.0) do not provide adequate search and navigation in the environment of distributed knowledge at the semantic level.

Naturally the thought came about some personal smart learning agents (software), which could identify relevant information from any accessible data source and provide an information synthesis tailored to personal learning objective. The idea of semantic educational portals that could provide a meaningful integration of educational objects with the adaptation and personalisation of training courses and curricula, appeared almost simultaneously with the advent of the Semantic Web.

During the semantization the data are combined into triplets in accordance with the RDF model and form a graph. If the data are the learning objects, than that form the so-called knowledge graph. It is obvious that the most adequate repository for the knowledge graphs are the graph databases.

The semantic graph database, also referred to as an RDF triplestore, stands out from the other types of graph databases due to the possibility to support ontologies. The semantic graph database is capable to integrate heterogeneous data from many sources and create relationships between datasets. That database focuses on the relationships between entities and is able to infer new knowledge out of existing information. It is a powerful tool to use in relationship analytics and knowledge discovery.

2 Related Work and Novelty

A recent authoritative overview [14] deals with the Graph and RDF databases makes it possible to navigate among modern products and solutions in the field of the Semantic Web, where the leaders are AllegroGraph, ArangoDB, BlazeGraph, Cray, DataStax, Ontotext GraphDB, IBM Graph, MarkLogic, OrientDB, Neo4j, Stardog, Teradata, Aster, Virtuoso.

It looks very promising the cooperative project Ontotext and Impelsys on the joint using of the platforms GraphDB and Dynamic Semantic Publishing for the development of personalized adaptive learning.

The pilot project [7] which is considered in this article is based on the cloud semantic platform and uses network RESTful services. The preferred repositories for learning objects themselves are the remote data storages. The predecessor of this project is the Cloud cabinet of the Educational portal "Department online" [2]. The project under consideration has been implemented in the educational practice of National Research Nuclear University MEPhI, Russia.

RDF browser is the main highlight of the Semantic Educational Web Portal [7], which distinguishes it from most of the known solutions in the field of the Semantic Web. Once being in the desired place of the knowledge graph via the corresponding widget, then you can to perform a visual navigation in this graph, simply walking along its nodes.

There is a possibility to make a visual walk through the knowledge graph as far as you want in any direction, scooping up the data that appears. By focusing on a specific graph node, it is possible to obtain text metadata, media content and hypertext links that are associated with this node. Along with that the nearest neighborhood of the node becomes visible and accessible for navigation.

3 The Ontology

The fundamental technologies of the Semantic Web, the knowledge graphs for example, are based on a set of universal standards, as set down by the World Wide Web Consortium (W3C) international community [17]. From

Proceedings of the XIX International Conference "Data Analytics and Management in Data Intensive Domains" (DAMDID/RCDL'2017), Moscow, Russia, October 10–13, 2017

the point of view of semantic technology, the key standards that apply are the Resource Description Framework (RDF) and OWL (Web Ontology Language).

RDF(S) [18], or triplets, is the format uses to store data in knowledge graphs. OWL [19] is based on the Description Logics language which is designed to show the data schema and to represent rich and complex knowledge about hierarchies of things and the relations between things. It is complementary to RDF and allows for formalizing a data schema/ontology in a given domain of knowlrdge, separately from the data itself.

In the general case an ontology is a formal specification that provides sharable and reusable knowledge representation. An ontology includes descriptions of concepts and properties in a concrete domain of knowlrdge, relationships between concepts, constraints on how the relationships can be used and occasionally individuals as instances of concepts.

Figure 1 partially shows the ontology [11] that is used in the Semantic Educational Web Portal [7]. In Figure 1, the Training_Course class is intentionally highlighted, because later this class and its individuals will be used as explanatory examples.

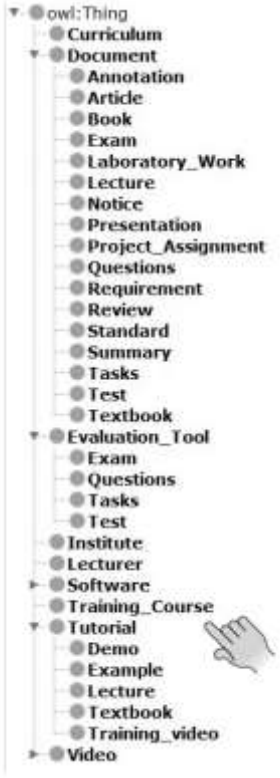


Figure 1 The class hierarchy in the ontology

Often, ontologies are understood as special knowledge repositories that can be read and understood both by people and computers, alienated from the developer and reused. Ontology in the context of information technology is usually a hierarchical system of concepts and terms (structure, model) of a certain subject area. Informally, an ontology is a description of the world view as applied to a particular area of interest. This description consists of terms and rules for the use of these terms, limiting their meaning within a particular area. At the formal level, an ontology is a hierarchical system consisting of a set of concepts and a set of assertions about these concepts on the basis of which it is possible to describe classes, relations, functions, and individuals (instances of classes).

In the language of Description Logics (DL) [4], a set of assertions of a general kind, or terminology, is called TBox (intensional knowledge). It is TBox that forms an ontology in the proper sense of the word. In Description Logics, sets of assertions of a individual kind – ABox (extensional knowledge) are singled out separately. TBox together with ABox forms a meaningful knowledge base (knowledge graph).

Below in Figure 2 is an example of the relationship between the class and individuals. Here the individual named Semantic_Web belongs to the class named Training_Course. In addition, this individual has a number of relations with individuals of other classes. This can be a relations of different types and directions, as can be seen from the color and direction of the arrows in Figure 2.

The very kinds of relations, like classes, are usually defined in TBox, whereas the facts of the existence of a certain kind of relationship between concrete individuals are intrinsically some RDF-assertion in ABox and each assertion has a triplet appearance.

Below Figure 3 shows a diagram of the relationship between classes from the ontology. This diagram presents only the top-level relationships. Every beam of particular color is a set of relations between individual instances of two classes.

Each individual relation in the ontology (that is in the knowledge graph) inherently is an RDF assertion where the subject is an instance of one class, the object is an instance of another class, and the reference is a predicate in the RDF format.

Depending on the number of relations between instances of two classes, every beam on diagram in Figure 3 can be thicker or thinner and gets a color of the class with a large number of incoming relations. These relations can be in both directions (incoming, outcoming). The number of relations (links) between classes from the ontology is shown in the legend on the diagram in Figure 3.

4 Knowledge Graphs

An ontology enriched with extensional knowledge from a specific subject area is also called the knowledge graph or knowledge base. Extensional knowledge forms the contents of ABox. Practically, knowledge graphs are deployed in the graph database or in a different semantic repository (triplestore or RDF store).

Specifically, the Semantic Educational Web Portal [7] is located on the Ontotext S4 GraphDB cloud platform [11] (physically on the Amazon Web Services – AWS cloud platform [1]).

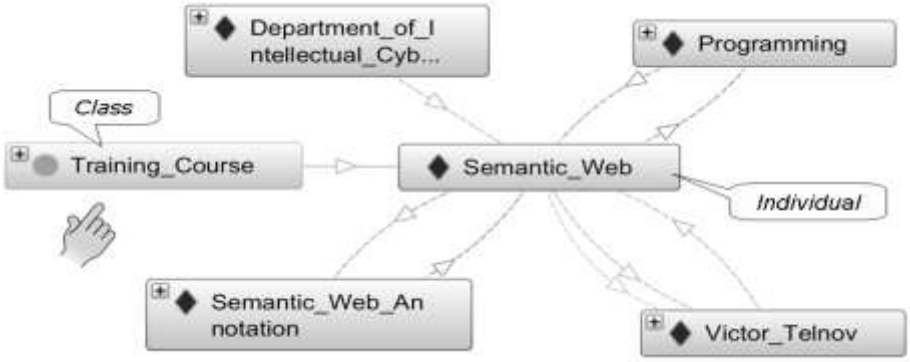


Figure 2 Individual of the class with neighborhood (example)

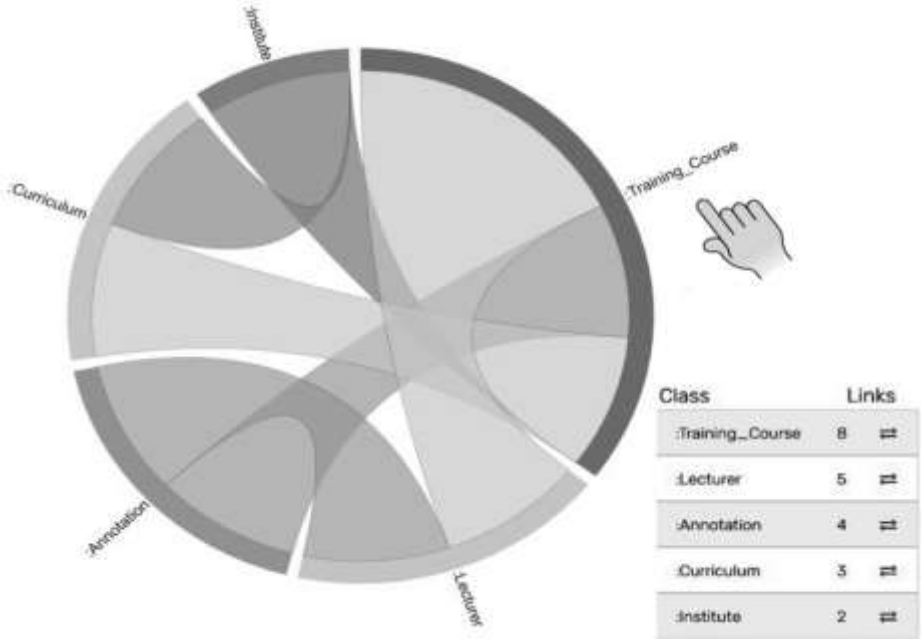


Figure 3 Relationships between classes in the ontology

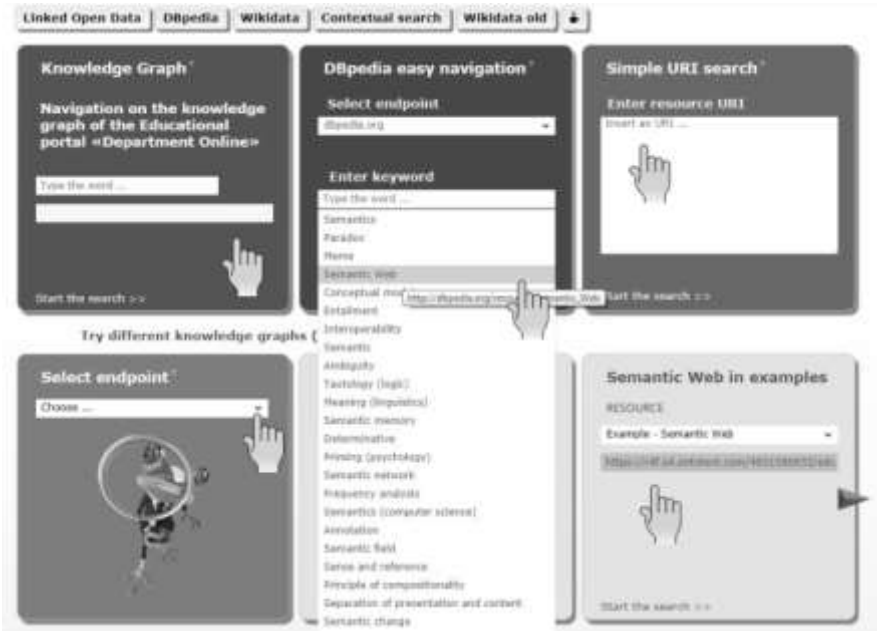


Figure 4 Navigation on the knowledge graphs

The current prototype of the Semantic Educational Web Portal [7] supports the curriculum presented in the Cloud cabinet of the Educational portal «Department online» [2]. Remote work with cloud version Ontotext GraphDB is carried out through the provided REST API. The most common operations are creating, reading, loading, and deleting semantic data. For the practical implementation of network requests HTTP methods are used, such as GET, POST, PUT, DELETE. These network requests contains essentially automatically generated SPARQL queries of the following types.

- SELECT to fetch data from the knowledge graph.
- CONSTRUCT to create a new RDF graph.
- INSERT to add triples to a graph.
- DELETE to remove triples from a graph.

Figure 4 shows the user interface of the Semantic Educational Web Portal, suitable for navigating through the available knowledge graphs. Primary graph of knowledge contains the materials of the master's courses, which are taught at the NRNU MEPhI on the profile "Computer networks and telecommunications". It is supplemented by the international knowledge bases DBpedia and Wikidata, as well as a number of more specialized knowledge repositories.

Each of mentioned knowledge graphs contains tons of triplets. The widgets shown below in Figure 4 are designed to allow a student or teacher easily get into the right place of the right knowledge graph, where it is likely find the required learning objects.

The principle of working with these widgets is largely similar to how information is searched through popular public search engines (Google, Yandex, etc.). As the user types the letters of the keyword in the input line, the system rolls out a list of relevant concepts from the knowledge graph. User can choose the most suitable concept and dive directly into the desired area of the graph.

Then, more accurate visual navigation on the knowledge graph becomes possible, which is performed in an intuitively clear manner using the RDF browser, as described below.

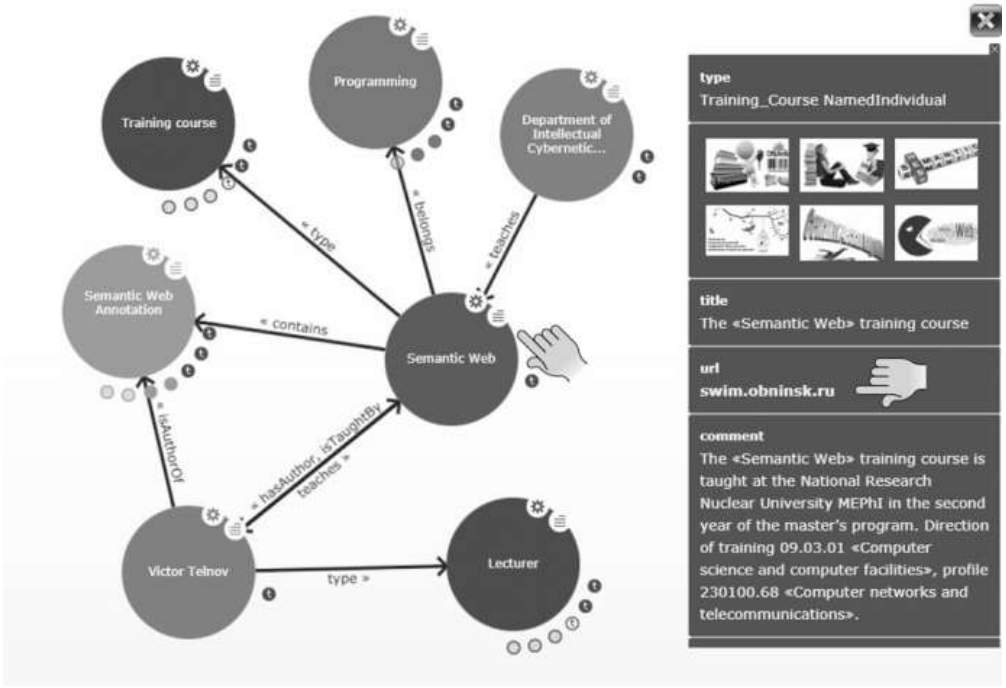


Figure 5 Fragment of the knowledge graph in the RDF browser (example)

5 RDF browser

RDF browser is the main highlight of the Semantic Educational Web Portal, which distinguishes it from most of the known solutions in the field of the Semantic Web. Having got to the right place of the necessary graph of knowledge through the corresponding widget, then you can perform a visual navigation in this graph, simply walking along its nodes.

By focusing on a specific graph node, it is possible to obtain text metadata, media content and hypertext links that are associated with this node. It is very important that the nearest neighborhood of the node becomes visible and accessible for navigation. This environment includes nodes not only of that graph, through which you originally has come in the semantic web, but also the nodes of all other knowledge graphs of that are supported by the system.

In Figure 5, some elements of the node's neighborhood that correspond to the Semantic_Web individual are displayed, as well as some related metadata. If you focus on the next node that is displayed by the RDF browser, it also becomes available with its neighborhood and metadata.

Thus you can to make a visual walk through the graph of knowledge as long as you like in any direction, scooping up the data that appears. In Figure 5, this is not shown, but in reality, when you hover over different sections of a particular node, pop-up menus, additional information and prompts for various options for continuing navigation through the knowledge graph becomes availible.

6 Adaptive Learning Technology

The main challenge of e-learning systems is to provide training courses tailored to different students with different learning rate and knowledge degree. Adaptive learning technologies are based on the fact that each student is unique, learns at varying rates and comes with different levels of knowledge. Traditional methodology of instruction may force the student down a learning path that is either too elementary, resulting in lack of interest or too heavy to grasp the nuances of the course. Adaptive learning, aided by semantic technologies [8], considers learner's interaction with courses and assessment modules to create personalized learning paths.

The adaptive learning system generally includes the following three subsystems.

1. The subsystem of forming a model of the learner (student model).
2. Learning planning subsystem (instructional model).
3. A subsystem for evaluating training outcomes.

For the student model the most popular means of determining a student's skill level is the method employed in CAT (computerized adaptive testing). In Semantic Educational Web Portal «Department online» [5] various, not just computerized means for measuring a student's skill level are used. In fact, the same training course should be built in different ways, depending not only on the level of knowledge and abilities of students, but also on the learning objectives. For example, a training course in programming will look different for students who concentrate in the field of business informatics and in the field of computer networks.

To build the actual instructional model and to fill it with learning objects, the Ontotext S4 Text Analytics RESTful service [13] is actively used. The purpose of text analysis is to create sets of structured data (machine-readable facts) out of heaps of unstructured, heterogeneous documents.Text analytics involves a set of techniques and approaches towards bringing various textual content to a point where it is represented as data and then mined for insights/trends/patterns. Contextual authoring provides lecturers with related texts, images and concepts which enhance the training course, reduces the time and costs of authoring and editing new learning content. Automated content enrichment improves the quality of curriculum and allows for continuous authoring without interruption.

When constructing adapted training courses, they are usually optimized according to two criteria: the effectiveness and adaptability of training. From the mathematical point of view, in the idealized case the problem can be reduced to finding the shortest path in the knowledge graph. This question has been studied, for example, in [9]. Neo4j Graph Database [10] has built-in means for calculating the shortest paths in the graph.

In real educational practice, the process of constructing a specific training course in the process of formation of a curriculum largely is empirical procedure, based on the experience and knowledge of the lecturer. In order to assess the training outcomes and learning efficiency, the evaluation tools from the Cloud cabinet of the Educational portal "Department online" are used, see [2].

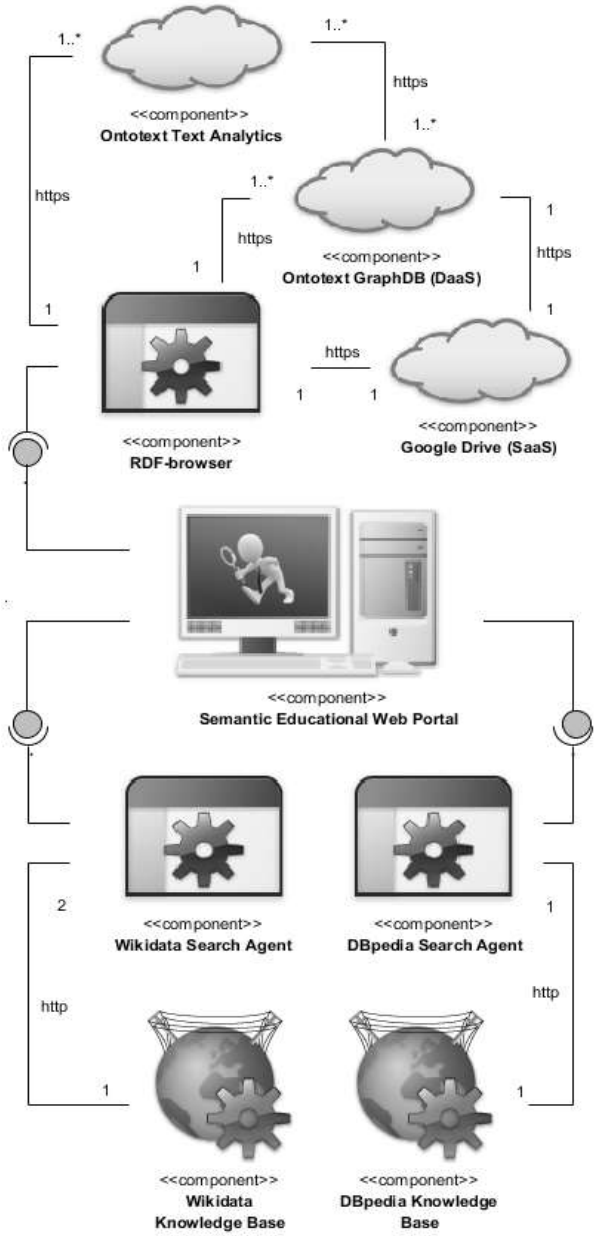


Figure 6 Software architecture

7 Software Architecture

Figure 6 shows the Deployment Diagram for the Semantic Educational Web Portal, performed in accordance with the UML 2 standard [6]. This diagram

can also be considered as an enlarged Component Diagram for this software. As it can be seen from Figure 6, the component named “RDF Browser” does not have its own server code (back-end). It interacts with two cloud RESTful services – Ontotext Text Analytics [13] and Ontotext GraphDB [11], both physically are deployed on the Amazon Web Services (AWS) [1] cloud platform.Cloud service Text Analytics [13] provides tools for semantic annotation and update of educational objects during the creation and adaptation of curricula. Cloud service GraphDB [11] provides the semantic storage for knowledge graphs and is mainly used as a SPARQL endpoint. As a universal repository for educational objects of an arbitrary nature, Google Drive is used. The choice of this particular storage is not principled, in parallel with it, arbitrary remote repositories equipped with data display means, for example such as Microsoft OneDrive or Yandex.Disk can be successfully applyed.

The other two components, named “Wikidata Search Agent” and “DBpedia Search Agent” both are advanced SPARQL endpoints to the corresponding international knowledge bases. Both mentioned components are provided with libraries of patterns of search queries, which largely facilitate the work of users, as well as are capable to deliver and show the found content in a variety of formats, including graphics.

8 Discussion

The pilot project presented in this article is aimed not only to provide students and teachers with a flexible knowledge management tool, but also to stimulate them to get acquainted with the world of semantic technologies.

To the middle of 2017 a sufficient toolkit for working with ontologies, knowledge graphs and semantic repositories of triplets, including on cloud platforms, has already been created. There is a great variety of public SPARQL endpoints. The English segment of the World Wide Web is filled with Linked Open Data. This is mainly reference data, bibliographic, media and other information of encyclopedic nature.

Attempts to find the open semantic data in the Russian segment of the World Wide Web infrequently lead to success. We have to agree with the fact, that in Russia there are still little Linked Open Data, suitable for educational activities. The main sources of data for Russian users of the semantic web are still international knowledge bases, including Russian-language content, primarily DBpedia [3] and Wikidata [15]. The prototype of the semantic educational web portal created is intended to partially fill this gap.

9 Concluding Remarks

A well-known skepticism about the fact that semantic educational portals will soon become widespread in the university environment seems fair. The modern realities of higher education are such that the overwhelming number of students and teachers do not suspect the existence of the Semantic Web and Linked Open Data. They continue to use traditional Content Management Systems (CMS), also known as Learning Management Systems (LMS) or Virtual Learning Environments (VLE), which are built primarily on simple taxonomies and thesauruses.

Students and professors widely practice searching the information on the World Wide Web for keywords, using public search engines for this purpose. Tradition plays a significant role here, as well as the simplicity and high speed of the search query generation, in comparison with the search queries to the Semantic Web.

Despite the growing commercialization of the public search engines, it can be assumed with a great deal of certainty, that they, along with Wikipedia, will remain the most accessible “universal textbooks” for the foreseeable future for that numerous category of students who not always demand the quality and completeness of the training material. An exception to this situation could be students (undergraduates) of universities who specialize in computer science and informatics.

10 Acknowledgements

The work was supported by the NBO “Vladimir Potanin Charity Fund”, project No ГК160001360.

References

[1] Amazon Web Services (AWS) – Cloud Computing Services (2017). https://aws.amazon. com/

[2] Cloud cabinet of the Educational portal «Department online» (2017). http://cloud.obninsk. ru/

[3] DBpedia (2017). https://ru.wikipedia.org/wiki/ DBpedia

[4] Description Logics (2017). http://dl.kr.org/

[5] Educational portal “Department online” (2017). http://ksst.obninsk.ru/

[6] ISO 19505 UML Part 2 Superstructure (2012). https://drive.google.com/file/d/0B0jk0QU2E5q9NV IwMFNieGxOZVU

[7] Knowledge graph of the Educational portal “Department online” (2017). http://semantic.obninsk. ru/

[8] Learning Resource Metadata Initiative (2017). http://lrmi.dublincore.net/

[9] Marwah, Alian1, Riad, Jabri: A Shortest Adaptive Learning Path in eLearning Systems: Mathematical View. J. of American Science, 5 (6), pp. 32-42 (2009). doi:10.7537/marsjas050609.08

[10] Neo4j Graph Database (2017). https://neo4j.com/

[11] Ontology of the Semantic Educational Web Portal (2017). http://drive.google.com/file/d/ 0B0jk0QU2E5q9Y0x6bTJaOEpXLWM

[12] Ontotext S4 GraphDB (2017). http://docs.s4. ontotext.com/display/S4docs/Fully+Managed+ Database

[13] Ontotext S4 Text Analytics (2017). http://docs. s4.ontotext.com/display/S4docs/Text+Analytics

[14] Philip Howard: Graph and RDF Databases 2016. Market Report Paper by Bloor. http://www.bloorresearch.com/research/market-report/graph-and-rdf-databases-2016/

[15] Victor Telnov: Semantic Web and Search Agents for Russian Higher Education. A Pilot Project. CEUR Workshop Proc. 1536. Selected Papers of the XVII Int. Conf. on Data Analytics and Management in Data Intensive Domains (DAMDID/RCDL 2015). Obninsk, Russia, pp. 195-204. http://ceur-ws.org/Vol-1536/

[16] Wikidata (2017). http://www.wikidata.org

[17] W3C Semantic Web (2017). https://www.w3.org/standards/semanticweb/

[18] W3C RDF Schema 1.1 (2014). https://www.w3.org/TR/rdf-schema/

[19] W3C OWL 2 Web Ontology Language (2012). https://www.w3.org/TR/owl2-overview/

Проблема транзитивности в системе категорий Википедии

© А.В. Кириллович

Казанский (Приволжский) федеральный университет,
Казань, Россия

alik.kirillovich@gmail.com

Аннотация. Проведен анализ системы категорий Википедии. Показано, что в системе категорий происходит нарушение требования транзитивности, в результате чего статья из подкатегории некоторой категории может не относиться к основной категории. С помощью стандартных методов онтологического моделирования проанализированы причины нарушений транзитивности. Предложен подход к автоматическому устранению этих нарушений.

Ключевые слова: Википедия, система категорий, информационно-поисковой тезаурус, онтологическое моделирование, OntoClean.

Problem of Transitivity of Wikipedia's Category System

© Alexander Kirillovich

Kazan (Volga Region) Federal University,
Kazan, Russia

alik.kirillovich@gmail.com

Abstract. This paper analyses a violation of the transitivity principle of Wikipedia's category system. Causes of the violation have been analyzed on base of ontological modeling methodologies such as OntoClean. A new approach for elimination of the violation has been proposed.

Keywords: Wikipedia, categorization system, thesaurus, ontology engineering, OntoClean.

1 Введение

Википедия – одно из крупнейших хранилищ информации. Данные Википедии используются в таких задачах? как разрешение лексической многозначности, категоризация текста, вычисление семантической близости, машинный перевод. Для автоматической обработки информации из Википедии требуется средство ее структурирования.

Система категорий – основной инструмент структурирования информации в Википедии. Категории хранят статьи, объединенные общей темой. Категории бывают двух видов:

- Категории-множества, например? Category:Cities, которая содержит статьи про конкретные города (Москва, Пекин, Лондон и т. д.).
- Категории-топики: например? Category:City, которая содержит статьи, относящиеся к городской тематике (*Городское планирование*, *Урбанизация*, *История городов* и т. д.).

Каждая категория может содержать подкатегории и самой находиться в родительских категориях. Категории могут группироваться с использованием мета-категорий, например,

Category:Writers→Category:Writers_by_nationality→Category:Russian_writers.

Таким образом, система категорий представляет собой ориентированный граф без циклов.

Согласно правилам Википедии, статья должна находиться в наиболее специфичной категории в иерархии. Например: статья Anton_Chekhov должна находиться не в категории Category:Writers, а в ее подкатегории Category:Writers→Category:Writers_by_nationality→Category:Russian_writers→Category:Russian_novelists→Anton_Chekhov. Поэтому, чтобы получить все статьи, релевантные заданной категории, требуется извлекать статьи и из всех вложенных подкатегорий. В связи с этим требуется, чтобы система категорий была транзитивной: статьи из вложенных подкатегорий должны быть релевантны родительской категории. Однако требование транзитивности нарушается, например:

- категория *Динамические системы* содержит вложенную подкатегорию *Социальные сети для ЛГБТ*: Dynamical_systems→Self-organization→Social_networks→Social_networking_services→LGBT_online_social_networking;

Труды XIX Международной конференции «Аналитика и управление данными в областях с интенсивным использованием данных» (DAMDID/ RCDL'2017), Москва, Россия, 10–13 октября 2017 года

- категория *Арифметика* содержит вложенную подкатегорию *Расстрелянные колумбийские революционеры*: Arithmetic→Ratios→Rates → Temporal_rates→Acceleration→Force→ Motion_ (physics)→Flight→Ballistics→...→ Projectile_ weapons→Firearms→People_associated_ with_firearms→Shooting_victims→...→ Colombian_people_executed_by_firing_squad.

Цель данной статьи – проанализировать проблему нарушения транзитивности в системе категорий Википедии и предложить подход к ее решению. Она построена следующим образом. В разделе 2 кратко перечислены другие средства структурирования данных Википедии и отмечены их ограничения. В разделе 3 исследована система категорий с помощью классических методов онтологического моделирования и выявлены причины потери транзитивности. В разделе 4 предложен подход, который исправляет систему категорий, удаляя из нее нетранзитивные цепочки. В разделе 5 приведена предварительная оценка предложенного метода на одном из разделов Википедии. В Разделе 6 описаны направления будущей работы.

2 Связанные работы

Существует множество проектов извлечения структурированных данных из Википедии [1, 2]: DBpedia [3–5], YAGO [6–8], WikiTaxonomy [9–11], WikiNet [12–14], ORA: The Natural Ontology of Wikipedia [15,16], WiBi [17], MENTA [18], BabelNet [19-20], WiSiNet [22], KOG: Kylin Ontology Generator [23], а также проекты выравнивания системы категорий и WordNet [24–28]. Однако данные проекты не являются универсальными и не применимы к категориям-топикам.

3 Анализ причин нарушения транзитивности

Проанализируем систему категорий с помощью методологий онтологического моделирования и выявим причины потери транзитивности.

Систему категорий можно рассмотреть в качестве тезауруса [29, 30]. Категории будут соответствовать концептам, а отношения между категорией и подкатегорией – стандартным отношениям между концептами онтологии:

- Отношение Класс→ Подкласс:
  - Cities_in_Europe→Category:Capitals_in_Euro pe
  - Category:Software → ... → Category: Operating_systems
  - Category:Mathematical_axioms → Category: Axioms_of_set_theory
  - Category:Machines → Category:Engines
  - Category:Wars → ... → Category:Wars_ involving_the_Soviet_Union
  - Category:Fiction_books → ... → Category:Dystopian_novels
- Отношение Класс→ Экземпляр:
  - Category:Capitals_in_Europe → Category:Moscow
  - Category:Intergovernmental_organizations → Category:United_Nations
  - Category:Universities_and_colleges_in_Conne cticut → Category:Yale_University
  - Category:Operating_systems→Category:Unix
  - Category:Fields_of_mathematics → Category:Algebra
  - Category:Axioms_of_set_theory → Category:Axiom_of_choice
  - Category:Abstract_strategy_games → Category:Chess
  - Category:Engines→Category:Internal_combus tion_engine
  - Category:Wars_involving_the_Soviet_Union →Category:World_War_II
  - Category:Dystopian_novels → Category:Nineteen_Eighty-Four
  - Category:Organs → Category:Brain
  - Category:Space_stations → Category:International_Space_Station
- Отношение Часть→Целое:
  - Category:Moscow → Category:Cities_and_towns_under_jurisdiction _of_Moscow → Category:Zelenograd
  - Category:Yale_University → Category:Yale_University_Library
  - Category:United_Nations → Category:International_Atomic_Energy_ Agency
  - Category:World_War_II → … → Category:Attack_on_Pearl_Harbor
  - Category:Central_nervous_system → Category:Brain
  - Category:Unix → Category:Network_socket
  - Category:Internal_combustion_engine → Category:Pistons
- Ассоциативные отношения:
  1. Наука → Объект изучения:
     - Category:Botany→Category:Plants
  2. Агент→Контрагент:
     - Category:Plants→Category:Herbicides
     - Category:Violence→ Category:Nonviolence
     - Category:Communism→ Category:Anti-communism
  3. Величина → Инструмент для измерения:
     - Category:Temperature→ Category:Thermometers

4. Деятельность→Агент деятельности:
    - Category:Hunting → Category:Hunting_dogs
    - Category:Military → Category:Military_personnel
5. Сырье→Результат:
    - Category:Grape→Category:Raisins
    - Category:Petroleum → Category:Petroleum_products→Gasoline
    - Category:Textiles→Category:Textile_arts →Category:Weaving
6. Другие ассоциативные отношения:
    - Category:Crops → Category:Crop_protection (Урожай → Защита урожая)
    - Category:Books→Category:Book_arts →Category:Bookbinding
    - Category:Death→ Category:Death_customs→ Category:Funerals
    - Category:Automobiles→ Category:Auto_racing

Мета-категориям соответствует так называемые «Node labels».

Представив систему категорий в виде тезауруса, мы применили к ней формальную методологию проверки корректности онтологий OntoClean, а также методологию построения информационно-поис-ковых тезаурусов [29–33]. В результате оказалось, что многие случаи нарушения транзитивности вызваны нарушениями правил построения иерархии концептов онтологии. Основными такими причинами являются:

- Неполное включение одной категории в другую. Примеры:
    - Англоязычный роман *«Лолита»* попал в категорию *Русские новеллы*: Category: Russian_novels→Category:Russian_novels_by_ writer→Category:Novels_by_Vladimir_ Nabokov→Lolita. Причина в том, что категория *Новеллы Набокова* не полностью входит в категорию *Русские Новеллы*.
    - *Японский язык* попал в категорию *Языки Кореи*: Category:Languages_of_Korea→ Category:Buyeo_languages→Category: Japonic_languages→Category:Japanese_language.
    - Аналоговая *Кинопленка* попала в категорию *Цифровые технологии*: Category:Digital_technology→Category: Digital_ media→Category:Video→Category:Film_ and_video_technology→Category:Film_stock;
- Ошибки при использовании нечетких понятий. Некоторые категории соответствуют понятиям с нечеткими границами. В результате вложенная глубоко категория может полностью выйти из-за пределов родительской. Например:
    - *Электрические стулья* попали в категорию *Потребительские товары*: Category: Consumer_goods→Category:Furniture→Category: Chairs→Category:Electric_chairs;
- Ошибки с использованием омонимичных категорий. Например:
    - *Музыкальные чарты* попали в *Диаграмы*: Diagrams→Charts→Record_charts. В одном случае Charts использовались в значении диаграмм, а в другом – в значении музыкальных чартов;
    - Строительство кораблей попало в Недвижимость: Real_estate→Construction→Ship_ construction;
- Использование одного понятия в разных смыслах. Например:
    - Электронная библиотека *Lib.ru* попала в категорию *Здания*: Buildings_and_structures → Buildings_and_structures_by_type→Libraries →Digital_libraries→Lib.ru. В одном случае *Библиотека* рассматривалась как тип здания, а в другом – как социальный институт;
    - Философское мировоззрение *Нигилизм* попало в категорию *Биология*: Category: Biology→Category:Life→Category:Philosophy_of_ life→Category:Nihilism. В одном случае *Жизнь* рассматривалась в значении биологический процесс, а другом случае – в значении социального процесса;
    - *Снег* попал в *Жидкости*: Category:Liquids → Category:Water→Category:Forms_of_water → Category:Snow. В одном случае *Вода* рассматривалась как вещество, а в другом – как это вещество в жидком состоянии;
- Несовместимые критерии идентичности. Например:

    Мусульманская святыня *Кааба* попала в категорию *Математических объектов*: Category:Mathematical_objects→Category: Geometric_shapes→Category:Elementary_shapes →Category:Cubes→Category:Cubic_buildings →Category:Kaaba. Ошибка находится в цепочке Category:Cubes→Category: Cubic_ buildings. Кубические здания, вообще говоря, не являются кубами, т. к. у них разные критерии идентичности. Куб – это абстрактный, вневременной, неизменный объект. Если куб изменится хотя-бы на миллиметр, то это будет уже другой куб. Кубическое здание – это конкретный объект, существующий во времени и пространстве и

сохраняющий идентичность при небольших модификациях;

- *Бермудский треугольник* попал в категорию *Геометрические объекты*: Category: Mathematical_objects → Category: Geometric_ shapes→Category:Elementary_shapes →Category:Triangles→Bermuda_Triangle;

- Смешение понятия и знака. Например:
  - *Династия Габсбургов* попала в категорию *Слова и фразы*: Words_and_phrases → ... → Surnames_of_Swiss_origin → Swiss_families →Swiss_noble_families→House_of_Habsburg
  - *Токсин* попал в категорию *Язык*. Language →Terminology→Biology_terminology→ Toxin. Причина ошибки в том, что Токсин не является термином. Термином является слово «Токсин»;
- Наследование типов от ролей. Например:
  - *Анальгетики* попали в категорию *Запрещенные лекарства*: Category: Illegal_drugs →Category:Morphine→Category:Analgesic. Причина ошибки в том, что Запрещенные лекарства – это не тип, а роль, и он не должен содержать категории-типы;
  - *Бомбовые прицелы* попали в категорию *Офисные принадлежности*: Category: Office_equipment→Category: Computers→ … →Category:Analog_computers→…→ Category:Optical_bombsights;
  - *Волчья ягода* (несъедобная) попала в категорию *Еда*: Category:Foods→Category: Fruit→Category:Berries→ Category: Sambucus.

В следующих случаях транзитивность нарушается не вследствие ошибки, а вследствие самого принципа построения системы категорий Википедии:

- Нетранзитивность отношения Класс→ экземпляр. Примеры:
  - Город *Зеленоград* попал в *Европейские столицы*: Category:Capitals_in_Europe → Category:Moscow→…→Category:Zelenograd;
  - Собака *Блонди* попала в категории *Нацистских лидеров*: Category:Nazi_leaders→ Category:Adolf_Hitler → Blondi;
  - *Атака на Перл-Харбор* попала в категорию *Войны СССР*: Category: Wars_involving_ the_Soviet_Union→Category:World_War_II →…→Category:Attack_on_ Pearl_Harbor;
  - Корабль *«Санта-Мария»* попал в категорию *Типы кораблей*: Category:Ship_types→...→ Category:Exploration_ships→Santa_María_ (ship);
  - Поэма под названием *«Ода»* попала в категорию *Жанры литературы*: Category: Literary_genres→Category:Poetry→...→ Ode_(poem). Этот пример особенно опасен, т. к. существует настоящий литературный жанр с таким именем.
- Нетранзитивность ассоциативного отношения. Примеры:
  - Вымышленная книга *«Теория и практика олигархического коллективизма»* из романа «1984», написанная, согласно сюжету романа, заведомо позже 1948 года, попала в категорию *Новеллы 1948 года*: Category: 1948_novels→Category:Nineteen_Eighty-Four→The_Theory_and_Practice_of_Oligarchical_Collectivism;
  - *Галактическая Империя* из вымышленной вселенной «Звездных войн» попала в категорию *Североамериканские государства*: Category:Northern_American_countries→ United_States→Category:American_people →…→Category:George_Lucas→Category: Star_Wars→…→Galactic_Empire_ (Star_Wars);
  - *Языки Джибути* попали в категорию *Статистика*: Category:Statistics→ Category: Statistical_data_sets→Category:Demographics_ by_country→Category:Demographics_of_Djibouti→Category:Languages_of_Djibouti;
  - *Биологическое оружие* попало в категорию *Трудовое право*: Category:Labour_law →Category:Labour_relations→Category: Occupational_safety_and_health→Category: Toxicology→Category:Biological_weapons;
  - *Расстрелянные колумбийские революционеры* попали в категорию *Арифметика*: Category:Arithmetic→Category:Ratios→Category:Rates→Category:Temporal_rates→Category: Acceleration→Category:Force→Category: Motion_(physics)→Category:Flight→ Category:Ballistics→…→Category:Projectile_ weapons→Category:Firearms→Category:People_associated_with_firearms→Category:Shooting_ victims→...→Category:Colombian_people_ executed_by_firing_squad;

Итак, нарушение транзитивности в системе категорий Википедии вызвано двумя группами причин. К первой группе относятся причины, связанные с нарушением правил построения иерархии концептов в онтологии. Эти нарушения могут быть устранены самими авторами Википедии (и действительно устраняются по мере развития проекта). Ко второй группе относятся причины, связанные с самим принципом устройства системы категорий Википедии, главная из которых – нетранзитивность ассоциативного отношения.

4 Подход к устранению нарушений транзитивности

Предложим метод, который устраняет не транзитивные цепочки и оставляет только транзитивные.

Как было показано в предыдущем разделе, одна из основных причин нарушения транзитивности состоит в том, что некоторые категории связаны с подкатегориями ассоциативным отношением, которое в общем случае не является транзитивным.

Существующие методы извлечения структурированной информации из системы категорий (например, YAGO или WikiTaxonomy) выявляют ассоциативные связи между категориями и просто устраняют их. Недостатком данных методов является то, что они исключают даже те ассоциативные связи, которые не нарушают транзитивность. В связи с этим возникает потребность в методе, который устраняет ассоциативные отношения, нарушающие транзитивность (например, Статистика → Демография), но сохраняет не нарушающие (например, Образование → Учитель). Опишем основные принципы этого метода.

Данный метод основан на подходе, который применяется в тезаурусе РуТез для установления ассоциативных отношений между концептами [30, 34–37]. В соответствии с этим подходом ассоциативное отношение между двумя концептами является транзитивным, если между концептами существует отношение онтологической зависимости.

Для формализации отношения онтологической зависимости [38–41] в РуТезе используется так называемый модально-экзистенциональный подход: объект А зависит от объекта В тогда и только тогда, когда необходимо, что если существует А, то существует и В [42, 43].

Модально-экзистенциональный подход имеет ряд преимуществ, среди которых – простота и математическая строгость. Недостатком этого подхода является то, что его применение требует участия человека. В связи с этим в исходном виде он не применим для решения поставленной нами задачи. Кроме того, модально-экзистенциональный подход был подвергнут критике с чисто онтологической точки зрения. К. Файн показал, что данный подход является слишком грубым приближением к понятию онтологической зависимости и имеет ряд контрпримеров [44, 45]. В качестве альтернативы Файн предложил эссенциональный подход к формализации онтологической зависимости. Согласно этому подходу, А зависит от В, если А является неустранимой компонентой сущности В. При этом сущность объекта понимается как утверждения, истинные в силу идентичности этого объекта. Эти утверждения, в свою очередь, образуют реальное определение объекта [44, 45].

Таким образом, использовав эссенциональный подход, мы получили следующий критерий определения онтологической зависимости: X онтологически зависит от Y, если Y неустранимым образом входит в определение X. Данный подход гораздо лучше подходит для автоматического применения. В качестве определения объекта, соответствующего той или иной категории, мы брали аннотации главной статьи этой категории и этой статьи на других языках. Факт вхождения объекта в определение другого объекта моделировался как наличие гиперссылки между определениями.

Таким образом, предложенный метод работает следующим образом:

- Определяем, является ли отношение между категорией и ее подкатегорией ассоциативным. Полагаем, что отношение является ассоциативным, если в нем участвует категория-топик. Тип категории определяем с помощью метода из проекта WikiTaxonomy [11].
- Если отношение является ассоциативным, то с помощью описанного выше критерия проверяем отношение онтологической зависимости между подкатегорией и категорией. Если зависимость имеется, то сохраняем отношение между категориями, если не имеется, то устраняем.
- Если отношение является не ассоциативным, а таксономическим, то используем уже существующий ресурс YAGO, содержащий очищенные таксономические отношения. В случае, если отношение присутствует в YAGO, сохраняем его и удаляем в противоположенном случае.

5 Оценка

Мы провели предварительное оценивание нашего метода на категории Mathematics. Для этого с помощью данного метода мы исключили из этой категории предположительно нерелевантные ей подкатегории. Список удаленных и оставленных подкатегорий был передан для ручной оценки. Задача асессора состояла в том, чтобы оценить, действительно ли оставленные категории релевантны основной категории и действительно ли удаленные – не релевантны. Результат оценивания представлен в Таблице 1.

Таблица 1 Результат предварительной оценки предложенного метода на категории Mathematics

Total	4281
True positives	2136
True negatives	650
False positives	1010
False negatives	485
Recall	0,814956
Precision	0,678957
F1 score	0,740766

6 Заключение

Мы проанализировали причины нарушения транзитивности в системе категорий Википедии и

предложили подход к их устранению, а также провели предварительное оценивание предложенного подхода на одной из категорий.

В дальнейшем мы планируем доработать данный подход. В частности, предполагается извлекать определения не только из главных страниц категории, а также использовать контекст-ссылки внутри текста определения. Такой подход позволит извлечь тезаурус из системы категорий Википедии. Этот тезаурус будет отличаться от аналогичных тем, что в нем будут присутствовать не только иерархические, но и произвольные ассоциативные отношения.

Благодарности

Работа выполнена за счет средств субсидии, выделенной Казанскому федеральному университету для выполнения государственного задания в сфере научной деятельности, проект 1.2368.2017/ПЧ.

Литература

[1] Medelyan, O., Milne, D., Legg, C., Witten, I.H.: Mining meaning from Wikipedia. Int. J. of Human-Computer Studies, 67 (9), pp. 716-754 (2009)

[2] Hovy, E., Navigli, R., Ponzetto, S.P.: Collaboratively Built Semi-Structured Content and Artificial Intelligence: The Story so Far. Artificial Intelligence, 194, pp. 2-27 (2013)

[3] Auer, S., Bizer, Ch., Kobilarov, G., Lehmann, J., Cyganiak, R., Ives, Z.: DBpedia: A Nucleus for a Web of Open Data. 6th Int. Semantic Web Conf. (ISWC'07/ASWC'07). pp. 722-735 (2007)

[4] Bizer, C., Lehmann, J., Kobilarov, G., Auer, S., Becker, Ch., Cyganiak R., Hellmann, S.: DBpedia: A Crystallization Point for the Web of Data. J. of Web Semantics, 7 (3), pp. 154-165 (2009)

[5] Lehmann, J., Isele, R., Jakob, M., Jentzsch, A., Kontokostas, D., Mendes, P.N., Hellmann, S., Morsey, M., van Kleef, P., Auer, S., Bizer Ch.: DBpedia: A Large-scale, Multilingual Knowledge Base Extracted from Wikipedia. Semantic Web J., 6 (2), pp. 167-195 (2015)

[6] Suchanek, F.M., Kasneci, G., Weikum, G.: Yago: a Core of Semantic Knowledge. 16th Int. Conf. on World Wide Web (WWW 2007), pp. 697-706 (2007)

[7] Hoffart, J, Suchanek, F.M., Berberich, K., Weikum, G.: YAGO2: A Spatially and Temporally Enhanced Knowledge Base From Wikipedia. Artificial Intelligence, 194, pp. 28-61 (2013)

[8] Mahdisoltani, F., Biega, J., Suchanek, F.M.: YAGO3: A Knowledge Base from Multilingual Wikipedias. 6th Conf. on Innovative Data Systems Research (CIDR 2015) (2015)

[9] Ponzetto, S.P., Strube, M.: Deriving a Large Scale Taxonomy from Wikipedia. 22nd National Conf. on Artificial Intelligence (AAAI 2007), pp. 1440-1445 (2007)

[10] Ponzetto, S.P., Strube, M.: Taxonomy Induction Based on a Collaboratively Built Knowledge Repository. Artificial Intelligence, 175 (9–10), pp. 1737-1756 (2011)

[11] Zirn C., Nastase V., Strube, M.: Distinguishing between Instances and Classes in the Wikipedia Taxonomy. 5th European Semantic Web Conf. (ESWC 2008), pp 376-387 (2008)

[12] Nastase V., Strube, M.: Decoding Wikipedia Categories for Knowledge Acquisition. 23rd National Conf. on Artificial Intelligence (AAAI 2008), pp. 1219-1224 (2008)

[13] Nastase V., Strube, M., Boerschinger, B., Zirn, C., Elghafari, A.: WikiNet: A Very Large Scale Multi-Lingual Concept Network. 7th Int. Conf. on Language Resources and Evaluation (LREC 2010), pp. 1015-1022 (2010)

[14] Nastase V., Strube, M.: Transforming Wikipedia into a Large Scale Multilingual Concept network. Artificial Intelligence, 194, pp. 62-85 (2013)

[15] Gangemi, A., Nuzzolese, A.G., Presutti, V., Draicchio, F., Musetti, A., Ciancarini, P.: Automatic Typing of DBpedia Entities. 11th Int. Semantic Web Conf. (ISWC 2012), pp. 65-81 (2012)

[16] Nuzzolese, A.G., Gangemi, A., Presutti, V., Ciancarini, P.: Towards the Natural Ontology of Wikipedia. 12th Int. Semantic Web Conf. (ISWC 2013). The Posters & Demos Track Proceedings., pp. 273-276 (2013)

[17] Flati, T., Vannella, D., Pasini, T., Navigli, R.: Two Is Bigger (and Better) Than One: the Wikipedia Bitaxonomy Project. 52nd Annual Meeting of the Association for Computational Linguistics (ACL 2014), pp. 945-955 (2014)

[18] de Melo G., Weikum, G.: MENTA: Inducing Multilingual Taxonomies from Wikipedia. 19th ACM Conf. on Information and Knowledge Management (CIKM 2010), pp. 1099-1108 (2010)

[19] Navigli, R., Ponzetto, S.P.: BabelNet: Building a Very Large Multilingual Semantic Network. 48th Annual Meeting of the Association for Computational Linguistics (ACL 2010), pp. 216-225 (2010)

[20] Navigli, R., Ponzetto, S.P.: BabelNet: The Automatic Construction, Evaluation and Application of a Wide-coverage Multilingual Semantic Network. Artificial Intelligence, 193, pp. 217-250 (2012)

[21] Moro A., Navigli, R.: WiSeNet: Building a Wikipedia-based Semantic Network with Ontologized Relations. 21th ACM Conf. on Information and Knowledge Management (CIKM 2012), pp. 1672-1676 (2012)

[22] Wu, F., Weld, D.S.: Automatically Refining the Wikipedia Infobox Ontology. 17th World Wide Web Conf. (WWW 2008), pp. 635-644 (2008)

[23] Ruiz-Casado M., Alfonseca E., Castells P.: Automatic Assignment of Wikipedia Encyclopedic Entries to WordNet Synsets. 3rd Int. Atlantic Web Intelligence Conf. (AWIC 2005), pp. 380-386 (2005).

[24] Toral, A., Muñoz, R., Monachini, M.: Named Entity WordNet. 6th Conf. on Language Resources and Evaluation (LREC 2008), pp. 741-747 (2008)

[25] Niemann, E., Gurevych, I.: The People's Web Meets Linguistic Knowledge: Automatic Sense Alignment of Wikipedia and WordNet. 9th Int. Conf. on Computational Semantics (IWCS 2011), pp. 205-214 (2011)

[26] Ponzetto, S.P., Navigli, R.: Large-Scale Taxonomy Mapping for Restructuring and Integrating Wikipedia. 21st Int. Joint Conf. on Artificial Intelligence (IJCAI 2009), pp. 2083-2088 (2009)

[27] Gella S., Strapparava C., Nastase V.: Mapping WordNet Domains, WordNet Topics and Wikipedia Categories to Generate Multilingual Domain Specific Resources. 9th Int. Conf. on Language Resources and Evaluation (LREC 2014), pp. 1117-1121 (2014)

[28] Titze, G., Bryl, V., Zirn, C., Ponzetto, S.P.: DBpedia Domains: Augmenting DBpedia with Domain Information. 9th Int. Conf. on Language Resources and Evaluation (LREC 2014), pp. 1438-1442 (2014)

[29] ANSI-NISO Z39.19-2005. http://www.niso.org/apps/group_public/download.php/12591/z39-19-2005r2010.pdf

[30] Loukachevitch, N.: Thesauri in Information Retrieval Tasks. Moscow University Press, 2011

[31] Guarino, N., Welty, Ch.: An Overview of OntoClean. Handbook on Ontologies (2nd edition). Springer (2009)

[32] Guarino, N., Welty, Ch.: A Formal Ontology of Properties. 12th European Workshop on Knowledge Acquisition (EKAW 2000) (2000)

[33] Guizzardi, G.: Ontological Foundations for Structural Conceptual Models. Telematica Instituut / CTIT (2005) https://research.utwente.nl/en/publications/ontological-foundations-for-structural-conceptual-models

[34] Loukachevitch, N., Dobrov, B.: RuThes Linguistic Ontology vs. Russian Wordnets. 7th Conf. on Global WordNet (GWC 2014), pp. 154-162 (2014)

[35] Loukachevitch, N., Dobrov, B., Chetviorkin, I.: RuThes-Lite, a Publicly Available Version of Thesauru of Russian Language RuThes. 20th Annual Int. Conf. "Dialogue", pp. 340-349 (2014)

[36] Loukachevitch, N., Dobrov, B.: Development of Ontologies with Minimal Set of Conceptual Relations. 4th Int. Conf. on Language Resources and Evaluation (LREC 2004), pp. 1889-1892 (2004)

[37] Loukachevitch, N., Dobrov, B.: Ontological Types of Associative Relations in Information-Retrieval Thesauri and Automatic Query Expansion. Ontologies and Lexical Resources in Distributed Environments (Ontolex 2004), pp. 24-29 (2004)

[38] Tahko, T.E., Lowe, E.J.: Ontological Dependence. Stanford Encyclopedia of Philosophy (2015). https://philpapers.org/rec/TAHOD. doi: 10.1080/05568649409506409

[39] Correia, F.: Ontological Dependence. Philosophy Compass, 3 (5), pp. 1013-1032 (2008)

[40] Koslicki, K.: Varieties of Ontological Dependence. Metaphysical Grounding: Understanding the Structure of Reality. Cambridge University Press, pp. 186-213 (2012)

[41] Koslicki, K.: Ontological Dependence: An Opinionated Survey. Varieties of Dependence. Benjamin Schnieder, Miguel Hoeltje and Alex Steinberg (eds.), Varieties of Dependence: Ontological Dependence, Grounding, Supervenience, Response-Dependence (Basic Philosophical Concepts). Philosophia Verlag, pp. 31-64 (2013). https://philpapers.org/rec/KOSODA-3

[42] Simons, P.: Parts: A Study in Ontology. Clarendon Press, Ch. 8. Ontological Dependence (1987). https://global.oup.com/academic/product/parts-9780199241460?cc=ru&lang=en&

[43] Thomasson, A.L.: Fiction and Metaphysics. Cambridge University Press. Chapter 2. The Nature and Varieties of Existential Dependence, pp. 24-34 (1999)

[44] Fine, K.: Essence and Modality. Philosophical Perspectives, (8), pp. 1-16 (1994). https://philpapers.org/rec/FINEAM-2

[45] Fine, K.: Ontological Dependence. Proc. of the Aristotelian Society, 95, pp. 269-290 (1995). https://philpapers.org/rec/FINOD

Подход к фильтрации запрещенного контента в веб-пространстве

© Е.А. Сидорова[1,2] © И.С. Кононенко[1,2] ©Ю.А. Загорулько[1,2]

[1] Институт систем информатики имени А.П. Ершова СО РАН,
[2] Новосибирский государственный университет,
Новосибирск, Россия

lsidorova@iis.nsk.su irina_k@cn.ru zagor@iis.nsk.su

Аннотация. Введение законодательного регулирования содержания информационных ресурсов обострило проблему автоматического обнаружения и блокировки запрещенного контента. Предложен подход к решению данной проблемы, в котором тематический анализ веб-сайтов дополняется жанровым, что позволяет выявить осуществляемую посредством веб-сайта деятельность и, благодаря этому, более точно распознать и локализовать запрещенный контент. Решение о наличии запрещенного контента на странице сайта принимается не только на основе анализа ее содержимого, но и на основе результатов анализа тематики и жанра сайта в целом. Разработаны программные средства и русскоязычные ресурсы для обнаружения запрещенного контента, относящегося к теме «Наркомания и наркотики».

Ключевые слова: классификация веб-сайтов, фильтрация запрещенного контента, тематический анализ текста, жанровый анализ веб-сайтов.

An Approach to Filtering Prohibited Content on the Web

© E.A. Sidorova[1,2] © I.S. Kononenko[1] © Yu.A. Zagorulko[1,2]

[1] A.P. Ershov Institute of Informatics Systems,
[2] Novosibirsk State University,
Novosibirsk, Russia

lsidorova@iis.nsk.su irina_k@cn.ru zagor@iis.nsk.su

Abstract. The institution of legislative regulation of the content of information resources has aggravated the problem of automatic detection and blocking of prohibited content. We propose an approach to solving this problem. In this approach, a thematic analysis of websites is complemented by a genre one, which allows identification of the activity carried out through a website and, therefore, brings about a more accurate recognition and localization of the illicit content. The decision on the presence of prohibited content on a website page is made on the basis of both analysis of the page text content and results of thematic and genre analysis of the site as a whole. Software and Russian-language resources for the detection of prohibited content related to the topic “Drug addiction and drugs” have been developed.

Keywords: website classification, filtering prohibited content, thematic text analysis, website genre analysis.

1 Введение

Задача избирательного распространения информации, сформулированная Луном (Luhn) в 1958 г., получила наименование «фильтрация» в 1975 г. (Denning). Система фильтрации контролирует поток документов, отбирая в нем полезные документы в соответствии с некоторым критерием (информационная потребность пользователя). Более полно задача определена в [5]: процесс фильтрации предназначен для отбора или удаления информации из динамического потока данных.

Введение законодательного регулирования содержания информационных ресурсов обострило проблему обнаружения и блокировки запрещенного контента, к которому относится любое запрещенное государством для просмотра и ознакомления информационное наполнение ресурса или веб-сайта (текст, мультимедиа, графика). При существующей скорости прироста и обновления информации в полной мере контролировать ее содержание с помощью модераторов-людей практически невозможно.

Современные подходы к автоматической фильтрации запрещенного контента чаще всего основаны на использовании списков ссылок на сайты (URL-фильтрация) [13], распознавании ключевых

Труды XIX Международной конференции «Аналитика и управление данными в областях с интенсивным использованием данных» (DAMDID/ RCDL’2017), Москва, Россия, 10–13 октября 2017 года

слов из списка запрещенных, а также на основе тематической классификации, например [6, 10]. Указанные методы не дают требуемого качества: в первом случае списки составляются вручную и не позволяют оценивать новые сайты, во-втором случае ключевые слова дают очень грубую оценку и либо ложно блокируют сайты с употреблением терминов в других смыслах, либо недостаточно полно покрывают способы выражения запрещенной информации. Что касается тематической классификации, то, помимо большой зависимости от обучающей выборки, она не позволяет определить цели, с которыми дается та или иная информация, что приводит к ложному срабатыванию фильтра, а для огромных массивов интернет-данных это недопустимо.

При рассмотрении различных методов фильтрации [3, 5], таких, как Boolean Information Filtering, Vector Space Model, Neural Networks и т. п., подчеркивается важность семантических проблем, т. е. проблем неоднозначности терминов (синонимия, полисемия, омонимия), затрудняющих сопоставление терминов в процессе содержательной фильтрации. Для преодоления семантических проблем, например, в [7], предложен метод, основанный на лингвистической онтологии, в качестве которой используется WordNet [2]. Основным недостатком такого подхода является трудоемкость построения лингвистической онтологии для заданного языка и предметной области.

В предлагаемом нами решении используется комплексный подход, при котором решение о запрещенности страницы принимается на основании не только ее тематики, но и прагматики, т. е. вида деятельности, осуществляемой посредством сайта в целом. Дополнение тематического анализа жанровым, а также использование лексических признаков, позволяющих явным образом задать семантику терминов, дает возможность более точно распознать и локализовать запрещенный контент.

2 Задача фильтрации контента

Фильтрация текстового контента традиционно рассматривается как разновидность информационного поиска. С другой стороны, фильтрацию можно рассматривать как особый случай классификации по двум категориям (релевантные и нерелевантные). В обзоре [4] сформулированы сходства и различия информационного поиска, фильтрации и бинарной категоризации. Фильтрация, в отличие от поиска, основана не на запросах, а на представлении индивидуальных или групповых интересов (профиль пользователя). Запрос – сиюминутный интерес, а профиль – долговременный (возможно меняющийся) интерес.

Базовое сходство всех направлений заключается в наличии следующих компонентов:

1. Представление веб-объекта (документа).
2. Представление информационного класса (информационной потребности, категории, профиля пользователя).
3. Сопоставление документа и класса с помощью алгоритмов, вычисляющих меру сходства.

Запрещенный контент − это любое содержательное наполнение веб-сайта, предоставление которого для просмотра и ознакомления запрещено государством. На территории РФ действует федеральный закон № 149-ФЗ «Об информации, информационных технологиях и о защите информации», в соответствии с которым устанавливаются основания для включения сайтов в список запрещённых. Список тематик блокируемых ресурсов открыт и включает, к примеру, такие типы запрещенного контента, как: контент, предназначенный только для взрослых, пропаганда против отдельного лица, группы или организации; материалы, связанные с наркотиками; контент, связанный с оружием, и др.

Для апробации предлагаемого подхода в качестве запрещенного рассматривался текстовый контент на русском языке, относящийся к теме «Наркомания и наркотики».

В силу высокой сложности задачи выявления запрещенного контента предложенное решение основано на совокупности различных методов анализа текстов и интернет-документов, включая методы машинного обучения и инженерный подход.

Машинное обучение не является полностью автоматическим, оно также требует экспертной деятельности по аннотированию обучающего множества текстов метками классов. Однако сформированные автоматически (хотя и на основе экспертной разметки) описания классов содержат много «шумящей» лексики, которая на этапе классификации текстов понижает точность работы алгоритма.

Инженерный подход предполагает создание описаний классов с участием эксперта, который, используя ускоряющие его деятельность программные модули нормализации текста и генерации частотных словарей, формирует ресурсы для классификатора. Несмотря на трудоемкость реализации, инженерный подход обеспечивает высокое качество классификации текстов за счет экспертной фильтрации «шума» и дополнения словарей (описаний классов) недостающей лексикой, отсутствующей в обучающей коллекции.

Особенность предлагаемого решения состоит в интеграции тематических и жанровых методов классификации текстовых ресурсов на базе инженерных правил принятия решения о наличии вредоносного контента. Использование тематических градаций в теме «Наркомания и наркотики» обеспечивает построение ее описания во всем многообразии и полноту классификации контента.

Необходимость использования жанровой классификации вызвана особенностями основной

темы и требованиями к принимаемому решению – определению принадлежности контента к двум классам: запрещенному контенту и незапрещенному. Определение жанра позволяет уточнить решение, полученное на базе тематической классификации. Этому же способствуют используемые логические правила принятия решения о запрещенности контента, построенные на основе результатов жанровой и тематической классификации.

В силу особенностей текстов исследуемой тематики традиционный алгоритм обработки текстов дополнен модулем анализа специальной тематической и стилистически окрашенной лексики – научная терминология, сленг наркоманов, обесценная лексика, жаргон интернет-пользователей, тематическая лексика на латинице и транслите.

Для оптимизации времени работы приложения алгоритм реализуется в два этапа:

1. предварительный анализ: установление наличия в тексте лексики, характерной для заданной тематики;
2. основной алгоритм: тематическая и жанровая классификации с принятием окончательного решения о запрещенности / незапрещенности контента.

Предусмотрена возможность обоснования полученных решений путем предоставления промежуточных результатов работы алгоритма фильтрации в понятной для конечного пользователя форме: найденной лексики, полученной уточненной тематики, жанра и используемых решающих правил.

3 Модель знаний

Предлагаемое нами решение основано на использовании лингвистических и предметных знаний и включает следующие ресурсы:

1. Рубрикаторы: тематический, жанровый (жанры интернет-текстов), прагматический (жанры сайтов) и лексический (признаки терминов).
2. Предметный словарь, включающий тематическую и жанровую лексику.
3. Жанровые шаблоны веб-текстов.
4. Прагматические модели веб-сайтов.
5. Решающие правила.

Рассмотрим их подробнее.

Тематический рубрикатор вводит уточняющие подтемы для базовой тематики «Наркомания и наркотики» и включает как запрещенные темы, так и незапрещенные (см. Рис. 1).

Назначение данного рубрикатора:

- отделить сайты по заданной тематике;
- дать объяснение пользователю, почему сайт заподозрен или отнесен к запрещенным.

Жанровый рубрикатор предназначен для классификации веб-страниц и веб-сайтов по жанрам, что в дальнейшем используется как для уточнения тематической классификации, так и для повышения качества фильтрации на основе правил.

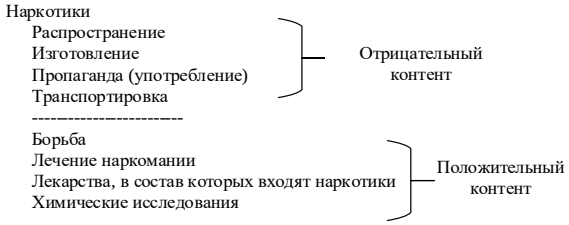


Рисунок 1 Фрагмент тематического рубрикатора

Выделяются такие жанры веб-ресурсов, как *Торговая площадка, Аптека, Сайт медицинской организации, Энциклопедический ресурс, Новостная лента, Персональная страница, Комментарий* и т. п.

Предметный словарь − структурированное хранилище терминов (слов и словокомплексов), в котором содержится вся необходимая информация для предварительного отбора тематически релевантных страниц, тематического и жанрового анализа текстового контента и принятия решения о блокировке.

Начальное наполнение словаря генерируется на этапе обучения с использованием размеченного экспертами корпуса веб-страниц, относящихся к исследуемой тематике, с применением универсального морфоанализатора, снабженного функцией предсказания незнакомой лексики.

Дополнительными источниками тематической лексики являются законодательно утверждённые Правительством РФ перечни наименований контролируемых наркотических средств, психотропных веществ и их прекурсоров, а также соответствующих видов растений, которые периодически пополняются и корректируются (примерно раз в год). Соответствующие документы доступны на официальных интернет-сайтах правовой информации, таких, как www.consultant.ru и pravo.gov.ru.

Далее осуществляется настройка предметного словаря экспертами, которые выделяют в его составе специальные подсловари, используя систему лексических признаков: тематическая лексика, научные термины, сленг наркоманов, термины на латинице, жанровая лексика и др. В задачу экспертов входит пополнение этих подсловарей, выявление регулярных ошибок фильтрации и формирование правил для изменения состава и структуры словаря.

Для создания и настройки словаря использовалась технология создания терминологических словарей KLAN [12].

Жанровые шаблоны веб-текстов формируются на основе лексических маркеров жанра и условий их встречаемости в текстовом фрагменте. Маркеры строятся на основе терминов словаря, при этом используются возможности представления совместной встречаемости терминов, альтернативности терминов в конкретной позиции (квазисинонимия), а также иерархической вложенности маркеров друг в друга. Например, страницы сайта типа *Торговая площадка* содержат следующие элементы:

- количественные конструкции (маркер: единица измерения “гр”, “мгр”),
- списки количественных конструкций (прайсы) с маркерами из жанровой лексики:

  Цены: 5гр. – 5 000 р, 10гр.— 9 000 р
- жанровая лексика: *цена, товар, закладка*.

Шаблон веб-страницы составляется из маркеров, на которые накладываются позиционные условия на тип фрагмента (заголовок, ссылка, выделенный текст, текст). Как и при описании маркеров, поддерживаются альтернативы и совместная встречаемость маркеров.

Рассмотрим для примера новостной шаблон:

«новостная лента»: [<_навигацияНовость, all_h>]
_навигацияНовость: [«главное за сутки»]
[«главное за сегодня»][«главное за день»]
[«все новости»][«основные новости»]
[«последние новости»][«лента новостей»]

Содержательно данный шаблон описывает следующее правило: если в одном из заголовков встретится один из маркеров группы *_навигацияНовость*, то это новостная лента.

Модель веб-сайта задается набором жанров веб-страниц, которые обязательно должны присутствовать на сайте и являются в совокупности его отличительным признаком. Для каждого сайта может быть задано несколько шаблонов. Например, модель интернет-магазина представлена двумя альтернативами:

[Магазин, Описание товара, ПредложениеТовара, Корзина, Доставка, Оплата]
[Магазин, Описание товара, ПредложениеТовара, СтатусЗаказа]

Принятие решения осуществляется на основе *решающих правил*, в посылках которых описываются условия того, будет ли анализируемый контент запрещен или разрешен. Эти условия строятся на термах, значениями которых являются конкретные тематики, жанры текста, жанры сайта и лексические признаки. Применяются правила двух видов: *положительные* и *отрицательные*, характеризующие текст, соответственно, как разрешенный или запрещенный. Правилами описываются, например, следующие экспертные наблюдения:

a) Если анализируемому контенту приписан лексический признак <40> «Обсценная лексика», он отнесен к тематике [601] «Употребление наркоманами» и жанру <401> «Торговая площадка» или (404) «Научная/информационная статья», то текст следует отнести к запрещенному контенту;

b) Текст по теме [1102] «Выращивание наркотических растений», написанный в жанре (407) «Словарная статья», относится к незапрещенному контенту. А текст по той же теме, представленный в ином жанре, может диагностироваться правилами как запрещенный контент и т. п.

Экспертные правила, помимо полноты, обладают высокой объяснительной способностью, что является существенным для нашей задачи.

Отметим, что правила принятия решений можно было бы сформировать автоматически при достаточном объеме обучающей выборки. Эксперимент показал, что экспертные правила не противоречат правилам, сформированным автоматически по обучающей выборке. Таким образом, можно рассматривать такой метод автоматического формирования правил как способ верификации правил, написанных экспертом.

4 Фильтрация контента

Анализ текстового контента осуществляется в несколько этапов. К основным этапам относятся тематическая и жанровая классификация текста, жанровый анализ сайта и принятие решения о запрещенности контента.

Объем статьи не позволяет в полной мере раскрыть каждый этап обработки текста, поэтому мы сконцентрируемся на основных идеях и используемых подходах.

4.1 Классификация текста

Прежде всего, необходимо уметь выявлять соответствие контента исследуемой тематике (*подозрительность* текста). При принятии решения о степени подозрительности контента необходимы:

а) Словарь тематической лексики, присутствие которой в тексте позволяет предположить тему «Наркомания и наркотики». Словарь содержит слова и словосочетания данного лексико-семанти-ческого поля, как специальные научные и нейтральные, так и жаргонные (сленг наркоманов). Эта лексика включает названия наркотиков, наркосодержащих лекарств и растений, названия состояний под воздействием наркотиков и т. п.

б) Критерий для определения возможной принадлежности к данной теме (степени подозрительности) текста, содержащего термины из словаря. Вычисление критерия опирается на степень присутствия тематической лексики с учетом лексического признака однозначности/ неоднозначности (омонимичная, т. е. тематически неоднозначная лексика из рассмотрения на данном шаге исключается).

Для подозрительных текстов применяется уточняющая классификация в соответствии с заданными рубриками с использованием весовых характеристик терминов, вычисляемых как ожидаемая взаимная информация (EMI) [9]. Данная мера позволяет оценить, сколько информации о классе – в теоретико-информационном смысле – содержит термин. Обучение и настройка алгоритма классификации производилась с участием эксперта.

При оценке релевантности текста классу (тематике) помимо веса термина учитывалась «зона текста», в которой встретился термин [1]: так, например, вес терминов в заголовках удваивался.

Способ взвешивания терминов, основанный на расчете EMI, дает улучшение на 5% по сравнению со способом взвешивания типа TF*IDF.

4.2 Жанровый анализ

В отличие от основной массы подходов к фильтрации, которые реализуют только контент-анализ страниц ресурсов, т. е. тематический анализ по ключевым словам, либо ограниченный жанровый анализ (преимущественно по формальным признакам, таким, как длина текста, количество букв, цифр и специальных признаков, количество ссылок и т. п. [8]), предложенный нами подход осуществляет многоаспектный жанрово-тематический анализ и классификацию. Используемые в рамках данного подхода признаки классификации явным или опосредованным образом отражают не только тематику анализируемых ресурсов, но и такие коммуникативно-прагматические аспекты жанра, как вид деятельности, осуществляемой посредством ресурса, включая цели и задачи деятельности и целевую аудиторию как ее участника, медийные свойства ресурсов, стилистические особенности используемых языковых средств.

Признаки жанрово-тематической классификации делятся на группы, каждая из которых отражает определенный аспект классификации:

1. Жанрово-структурная классификация ресурсов на основе двухуровневой модели:
   - Макроуровень – ресурс в целом;
   - Микроуровень (компоненты ресурса: страница, раздел, блок).
4. Жанрово-прагматическая классификация ресурсов (на основе прагматических аспектов содержания и представления):
   - Праксиологические (деятельностные) аспекты (вид деятельности, которая осуществляется посредством ресурса);
   - Аспекты содержания и представления, связанные с каналом коммуникации (медийные свойства ресурсов).
5. Жанрово-стилистическая классификация ресурсов:
   - Лексико-стилистические аспекты содержания и представления (стилистические особенности используемых языковых средств с акцентом на стилистически окрашенные языковые средства).

Представление о жанре закладывается на этапе формирования обучающей выборки, которая целенаправленно отбирается и размечается экспертами. Предлагаемая процедура жанровой классификации совмещает статистический и экспертный подходы к анализу жанра и опирается на метод вычисления меры принадлежности текста к жанру [11]. Вначале применяется экспертный подход, в рамках которого осуществляется поиск в тексте жанровых маркеров, т. е. сопоставление тексту шаблонов, составленных экспертом. Если на основе маркеров жанр веб-текста определить не удалось, то применяется классификация на основе методов машинного обучения.

4.3 Принятие решения на основе правил

Решение о запрещенности/незапрещенности контента принимается на основе следующих параметров:

1. $\bar{P}_t$=($p(t_1)$, $p(t_2)$, ..., $p(t_i)$, ..., $p(t_{Nt})$) – вектора релевантности текстового контента тематикам рубрикатора, где Nt – число тематик в рубрикаторе, $p(t_i)$ – вероятность реализации тематики t_i в анализируемом тексте, i = 1, ..., Nt; $\sum_{i=1}^{Nt} p(t_i) = 1$;
2. $\bar{P}_j$ =($p(j_1)$, ..., $p(j_{Nj})$) – вектора релевантности контента текста жанрам текста, заданным в жанровом рубрикаторе, где Nj – число жанров текста в рубрикаторе; $\sum_{i=1}^{Nj} p(j_i) = 1$;
3. $\bar{P}_{js}$ =($p(js_1)$, ... ,$p(js_{Ns})$) – вектора релевантности контента всего сайта жанрам сайта, заданным в рубрикаторе, где Ns – число жанров сайта в рубрикаторе; $\sum_{i=1}^{Ns} p(js_i) = 1$;
4. $VL=(v(lex_1), \ldots, v(lex_{Ln}))$ – вектора наличия лексических признаков в текстовом контенте, где $v(lex_i) \in \{0,1\}$ – показатель присутствия/ отсутствия в тексте лексического признака lex_i (например, сленга, обсценной лексики и т. п.);
5. $\bar{P}_{Rule}$ – набора решающих правил вида t_i & j_k & js_m & lex_j, принимающих решение о запрещенности / незапрещенности анализируемого контента в виде оценки m^p, вычисляемой как вероятность совместной реализации темы t_i, жанра текста j_k, жанра сайта js_m и лексического признака lex_j. в этом контенте. Оценка m^p вычисляется по формуле $p(t_i) \cdot p(j_k) \cdot p(js_m) \cdot v(lex_j)$, т. е. это произведение вероятностей указанных в правиле параметров, взятых из векторов, описанных выше;
6. $\bar{M} = (M^-, M^+)$ – двухкомпонентный вектор сумм оценок всех отрицательных и положительных правил соответственно.

Окончательное решение о запрещенности / незапрещенности контента принимается по критерию C: если $C = (M^- - M^+) > 0$, то считается, что контент запрещен. Настройка данного критерия позволяет изменять результаты работы системы в сторону повышения либо полноты, либо точности фильтрации.

5 Архитектура системы фильтрации запрещенного контента

Схема выявления запрещенного контента представлена на Рис. 2. На вход системы фильтрации запрещенного контента поступает контент сайта, представленный множеством веб-текстов (текстов с html-разметкой), либо обновление сайта – множество новых либо отредактированных веб-текстов сайта. Веб-текст – это единица текстового контента сайта, хранящаяся в БД на сервере. Веб-страница, которую видит пользователь при просмотре веб-сайта с

помощью веб-браузера на стороне клиента, формируется в общем случае из множества веб-текстов с добавлением незначащего для анализа контента – элементов оформления страницы, баннеров, рекламы и т. п., а также медиа-контента.

Обработка сайта начинается с анализа его структуры, затем формируется начальный индекс сайта (в случае обновления сайта индекс модифицируется), фиксируются зависимости между веб-текстами. После этого тексты сайта последовательно анализируются.

Рисунок 2 Схема выявления запрещенного контента

Каждый веб-текст очищается от html-разметки (значащие элементы разметки, такие, как заголовки, ссылки, выделение фрагмента стилем, сохраняются), осуществляется лингвистический анализ текста, обеспечивающий поиск в нем терминов словаря, и сбор статистической информации. Далее производится оценка тематической принадлежности текста к базовой теме «Наркомания и наркотики» – т. н. «оценка подозрительности» текста (текст считается подозрительным, если его контент соответствует базовой теме). В определении подозрительности участвует только однозначная лексика, наличие которой позволяет снять возможную тематическую неоднозначность текста. Для неподозрительных текстов дальнейшая оценка запрещенности не проводится, определяется лишь жанр текста, который заносится в индекс сайта.

Жанровая классификация позволяет определить жанр текста на основе словаря маркеров и структурного анализа текста в соответствии с разметкой. Если на основе маркеров и жанровых шаблонов жанр веб-текста определить не удалось, то применяется уточняющая классификация на основе методов машинного обучения.

Уточняющая классификация обеспечивает не только определение жанра текста, но и уточнение (конкретизацию) его тематики в соответствии с типами противоправных и разрешенных действий в рамках темы «Наркомания и наркотики». При уточняющей классификации используется обученный на размеченном корпусе текстов предметный словарь. Результатом уточняющей классификации являются векторы релевантности текста темам и жанрам, которые сохраняются в индексе сайта.

После первичной обработки всех веб-текстов сайта осуществляется анализ его жанра. Каждый жанр сайта описывается одной или несколькими моделями. Модель сайта фиксирует набор жанров текста, которые обязательно должны встретиться на сайте данного жанра. Данные модели составляются экспертами вручную на основе анализа структуры веб-сайтов обучающей коллекции. Вычисление оценки степени соответствия сайта какому-либо жанру осуществляется по моделям сайтов и оценкам, полученным для жанров веб-текстов сайта. Полученные оценки для жанра веб-сайта и составляющих его веб-текстов сохраняются в индексе сайта.

Принятие решения о запрещенности сайта осуществляется на основе решающих правил, которые применяются только для подозрительных текстов. Особенностью параметра подозрительности текста является то, что он «распространяется» на все связанные тексты (связи между текстами фиксируются структурой сайта и хранятся в индексе сайта). Поэтому на стадии предварительной обработки осуществляется поиск всех подозрительных текстов по связям и выполнение уточняющей классификации для тех из них, для которых она ранее не проводилась. Результатом применения правил к тексту является оценка запрещенности страницы.

Оценка запрещенности всего сайта определяется как максимум из оценок запрещенности по всем текстам сайта.

6 Результаты эксперимента

Для оценки качества фильтрации были сформированы одна обучающая и две тестовых коллекции, содержащие веб-тексты:

1. Обучающая коллекция, состоящая из 468 веб-текстов на русском языке, относящихся к теме «Наркомания и наркотики». Все тексты размечены экспертами. Разметка включает экспертную оценку запрещенности / незапрещенности контента, тематику, жанр веб-текста и жанр веб-сайта, на котором был размещен данный текст.

6. Тестовая коллекция веб-текстов, включающая около 123 тыс. русскоязычных веб-страниц, часть которых относится к теме «Наркомания и наркотики», но не содержит запрещенный контент. Коллекция собрана вручную на основе сайтов Яндекс-каталога (https://yandex.ru/yaca).

7. Тестовая коллекция веб-текстов, включающая 569 веб-текстов на русском языке, содержащих

запрещенный контент по теме «Наркомания и наркотики».

Полученные коллекции включают веб-тексты различных функциональных стилей – от нормативных и официальных документов до сообщений и комментариев на форумах и в социальных сетях, – что позволяет адекватно оценить качество фильтрации на всем многообразии интернет-жанров. К сожалению, в открытом доступе отсутствуют размеченные коллекции текстов по данной тематике, чем объясняется небольшой объем первой и третьей коллекций, которые создавались нашими экспертами вручную. Объем веб-текстов в коллекциях варьировался от 213 до 65655 Кб.

На основе обучающего корпуса текстов был построен словарь, который в дальнейшем был дополнен терминами из специализированных словарей. Словарь содержит более 50 тыс. терминов (без учета стоп-слов). Его общий количественный и качественный состав отражен в Таблице 1.

Таблица 1 Терминологический состав словаря

Лексем	Слово-комплексов	Подозрительных	Жанровых	Сленг
24175	26540	5349	1895	3161

Как видно из таблицы 1, ключевые слова для предварительного отбора текстов по теме («подозрительные», т. е. однозначные тематические термины) составляют десятую часть объема словаря.

Оценка качества классификации была дана в виде показателей полноты (R), точности (P) и F-меры. Рассматривалась бинарная классификация (1) и уточняющая тематическая классификация (2). Оба сравниваемых метода основаны на машинном обучении, но во втором случае используется расширенный набор тем, причем для каждой из них указано, является ли она запрещенной или нет.

Таблица 2 Сравнение методов классификации

	R	P	F-мера	Скорость
(1)	52,0%	65,4%	57,9%	~ 0,07 мс
(2)	72,6%	69,7%	71,1%	~ 0,10 мс

Как видно из Таблицы 2, использование уточняющего тематического рубрикатора, построенного по специальной ориентированной на задачу фильтрации методике, позволило улучшить показатели полноты и точности в сравнении с бинарной классификацией (когда контент сразу классифицируется на два класса – запрещенный и незапрещенный), соответственно, на 20% и 10%. Однако эти показатели все еще являются низкими.

Таблица 3 Оценка качества фильтрации

	Кол-во (страниц)	Правильных ответов (%)
Нейтральная коллекция	~ 123 тыс.	99.4%
Отрицательная коллекция	569	86.99%

В Таблице 3 приведены оценки работы системы фильтрации, в которой тематическая классификация сочетается с жанровой и применяются решающие правила (отметим, что результаты, полученные тематическим классификатором, использовались здесь в качестве промежуточных.)

Таким образом, ошибка первого рода составила 0,6%, ошибка второго рода – 13,01%.

Большая часть ошибок обоих типов связана с неполнотой словаря. Так, возможны существенные лакуны в подсловарях латиницы и транслита (например, отсутствуют названия наркотиков *25i-nbome, JWH, нбоме, дживиаш*). Не всегда в словаре учтена возможная лексическая или лексико-морфологическая неоднозначность (например, *доб.* может представлять в тексте наркотик или сокращение от *добавочный*).

Ложно-положительная оценка характерна для страниц, которые не проходят предварительный этап фильтрации ввиду отсутствия однозначной тематической лексики. Так, не блокируются (отсеиваются как неподозрительные) страницы, содержащие предложения или рекламу наркотических веществ, завуалированные путем использования неоднозначной лексики (например, *соли для ванн*), а также намеренно искаженные (зашифрованные) тексты.

Ложно-отрицательная оценка характерна для следующих типов веб-текстов: а) информационные статьи о наркотических веществах или растениях (в частности, о выращивании декоративных растений), жанр которых не определен как энциклопедическая/ словарная статья; б) новостные тематические тексты с позитивной окраской (*Умеренное потребление алкоголя и амфетамина может улучшить память у пожилых людей*); в) тематически нейтральные страницы комментариев на форумах и в блогах с вкраплением шутливых тематических комментариев (*Наркотой там не барыжите, случайно?* – реплика при обсуждении вопросов информационной безопасности).

Заключение

Предложенный подход реализован в виде приложения, интегрированного в платформу Plesk. Приложение позволяет выявлять и блокировать сайты, содержащие запрещенную информацию по теме «Наркомания и наркотики» и/или осуществляющие незаконную деятельность по торговле, распространению, транспортировке, изготовлению и пропаганде наркотиков.

К преимуществам предложенного подхода относятся, во-первых, глубокий анализ текстового контента веб-ресурса с учетом его тематических и жанровых особенностей, во-вторых, совмещение статистических и инженерных методов анализа текста, в частности, предложен уникальный метод принятия решения о запрещённости контента на основе решающих правил, учитывающих результаты его жанровой и тематической классификации, в-третьих, масштабируемость и технологичность разработанных программных средств, что позволяет

легко адаптироваться к различным предметным областям посредством настройки базы знаний.

В предложенном подходе, на наш взгляд, достигнут баланс между ручной работой эксперта и автоматическим обучением, где, во-первых, словари создаются и обучаются автоматически, а эксперты пополняют их номенклатурными терминами и сленгом, во-вторых, неполнота жанровых (функциональных) описаний интернет-ресурсов (создаются экспертом) компенсируется поддержкой статистического жанрового классификатора, и наконец, решающие правила потенциально могут строиться автоматически, а оценка применимости правила для каждого конкретного случая оценивается по вероятностной формуле.

Дальнейшее развитие описанной технологии связано с необходимостью автоматизации поддержки словаря в актуальном состоянии. Автоматизация возможна на базе жанрового анализа страниц, относящихся к жанрам «Нормативный список» (отслеживание словарей официальных наименований контролируемых веществ и растений) и «Словарная статья» (отслеживание словарей универсального и тематического сленга, обсценной лексики). Однако главным источником тематической лексики по-прежнему остаются эксперты, т. к. интернет-словари тематического сленга существенно отстают от происходящих в среде наркоманов изменений лексики.

В качестве актуального направления исследований по данной тематике также рассматривается возможность применения методов сентимент-анализа для улучшения распознавания трудноуловимой темы пропаганды наркотиков, представленной в информационных сообщениях, создающих привлекательный образ наркомана и процесса употребления наркотических веществ.

Благодарности

Работа выполнена при финансовой поддержке Министерства образования и науки Российской Федерации (договор № 02.G25.31.0054) и Российского фонда фундаментальных исследований (грант № 15-07-04144).

Литература

[1] Cohen, William W., Singer, Y.: Context-sensitive Learning Methods for Text Categorization. ACM Transactions on Information Systems, 17, pp. 141-173 (1999)

[2] Gormez, Josre M., Girarldez, I., De Buenaga, M.: Text Categorization for Internet Content Filtering. Inteligencia Artificial, Revista Iberoamericana de Inteligencia Artificial, 920, pp. 34-52 (2003)

[3] Khozooii, N.S., Haratizadeh, S., Keyvanpour, M.R.: An Analytical Framework for Web Information Filtering Techniques. Int. J. of Hybrid Information Technology, 6 (6), pp. 345-358 (2013)

[4] Nanas, N.: Literature Review: Information Filtering for Knowledge Management. The Open University, 2001. http://kmi.open.ac.uk/publications/ pdf/kmi-01-16.pdf

[5] Nouali O., Blache P. Automatic Classification and Filtering of Electronic Information: Knowledge-Based Filtering Approach. Int. Arab J. of Information Technology, 1 (1), pp. 85-92 (2004)

[6] Sebastiani, F.: Machine Learning in Automated Text Categorization. ACM Computing Surveys, 34 (1), pp. 1-47 (2002)

[7] Shoval, P., Maidel, V., Shapira, B.: An Ontology Content-based Filtering Method. Int. J. Information Theories & Applications, 15, pp. 303-314 (2008)

[8] Воронов, С.О., Воронцов, К.В.: Автоматическая фильтрация русскоязычного научного контента методами машинного обучения и тематического моделирования. Компьютерная лингвистика и интеллектуальные технологии: По материалам ежегодной Межд. конф. «Диалог». 2015. http://www.dialog-21.ru/digests/dialog2015/materials/pdf/VoronovSOVorontsovKV. pdf

[9] Маннинг, К.Д., Рагхаван, П., Шютце Х.: Введение в информационный поиск. М.: Вильяме, 528 с. (2011)

[10] Патент РФ № 2446460, МПК G06F21/20. Способ и система фильтрации веб-контента /Осипов Г.С., Тихомиров И.А., Соченков И.В.; патентообладатель ИСА РАН; заявл. 2010-11-18; опубл. 27.03.2012

[11] Сидорова, Е.А., Боровикова О.И.: Подход к жанровой классификации текстовых ресурсов. Информационные технологии и системы [Электронный ресурс]: Тр. Шестой Межд. науч. конф. ИТиС–2017: науч. электрон. изд. / отв. ред. Ю.С. Попков, А.В. Мельников. Челябинск: Челяб. гос. ун-т, сс. 264-269 (2017)

[12] Сидорова, Е.А.: Подход к построению предметных словарей по корпусу текстов. Труды межд. конф. «Корпусная лингвистика – 2008». СПб.: СПбУ, Факультет филол. и искусств, сс. 365-372 (2008)

[13] Стрекалов, И.Э., Новиков, А.А., Лопатин, Д.В.: Система формирования безопасности контента. Вестник ТГУ, 20 (2), сс. 462-464 (2015)

Специализированные инфраструктуры в ОИИД 1

Special-purpose DID infrastructures 1

Data Curation Policies for EUDAT Collaborative Data Infrastructure

© Vasily Bunakov[1], © Alexia de Casanove[2], © Pascal Dugénie[2], © Rene van Horik[3], © Simon Lambert[1], © Javier Quinteros[4], © Linda Reijnhoudt[3]

[1] Science and Technology Facilities Council, Harwell Campus, United Kingdom
[2] CINES, Montpellier, France
[3] Data Archiving and Networked Services (DANS), The Hague, Netherlands
[4] GFZ German Research Centre for Geoscience, Potsdam, Germany

vasily.bunakov@stfc.ac.uk, casanove@cines.fr, dugenie@cines.fr, rene.van.horik@dans.knaw.nl, simon.lambert@stfc.ac.uk, javier@gfz-potsdam.de, linda.reijnhoudt@dans.knaw.nl

Abstract. The work outlines an approach to the development of a data curation framework in the EUDAT Collaborative Data Infrastructure. Practical use cases are described as well as provisional results of defining granular data curation policies with high potential for their machine-executable implementation.

Keywords: data curation, e-infrastructures, long-term digital preservation, policies.

1 Introduction

EUDAT Collaborative Data Infrastructure (CDI) [1] is a European e-infrastructure of data services and information resources in support of research. This infrastructure and its services have been developed in close collaboration with over 50 research communities spanning across many different scientific disciplines, with more than 20 major European research organizations, data centres and computing centres involved. Researchers, research communities and service providers can use EUDAT data services to manage research data according to their own needs.

The EUDAT services offering has emerged as a result of two consecutive FP7 and Horizon 2020 projects, with the actual services focused on different aspects of data management and data use, and supported by a variety of information technology stacks. The major EUDAT services [19] are:

- B2ACCESS – identity and authorization service;
- B2HANDLE – service for assigning and managing persistent identifiers;
- B2DROP – service for secure and trusted data exchange;
- B2SHARE – service for sharing small-scale "long tail" data;
- B2SAFE – robust, safe and highly available service for storing large-scale data in community and departmental repositories;
- B2STAGE – service for managing data transfers between EUDAT storage and high-performance computing;
- B2FIND – service for data discovery across the EUDAT infrastructure (data catalogue).

Data curation (or digital curation) is the selection, preservation, maintenance, collection and archiving of digital assets and hence is the essential part of research data management. Sensible data curation requires establishing and developing long-term repositories of digital assets for their current and future use by researchers and wider society. Collaborative data infrastructures like EUDAT that span across the borders should play a significant role in research data curation.

Historically, EUDAT services have been built with only a few considerations for conscious data curation, with secure and controlled access to data being one of the major initial goals to achieve. Other aspects of data curation started playing a more prominent role when services matured to production stage and became a part of an operational collaborative infrastructure. Specifically, operational requirements of B2SAFE service (that currently offers what long-term digital preservation projects typically call "bit-level" preservation), as well as automated data transfers across interrelated B2DROP, B2SHARE and B2FIND services have made it essential to systematically explore the topic of data curation in EUDAT.

Proceedings of the XIX International Conference "Data Analytics and Management in Data Intensive Domains" (DAMDID/RCDL'2017), Moscow, Russia, October 10–13, 2017

The decision was made to formulate the core approach to data curation with the involvement of two prominent unrelated research communities with substantial amounts of data to manage and then, using these two use cases as a proof-of-concept for clearly formulated data curation activities, get other user communities involved.

Another decision made was to reuse the outputs of the SCAPE project [2] and Research Data Alliance Practical Policy Working Group [3] in order to set up a reasonable data curation framework for EUDAT.

The rest of the paper outlines the core use cases, characterizes the SCAPE and RDA outputs that are deemed to be applicable in EUDAT context, describes mapping of SCAPE policy elements [4] to granular data policies in EUDAT, and sets directions for further works on data policies in EUDAT.

2 HERBADROP use case

2.1 Motivations and relation to EUDAT services

The HERBADROP data pilot [12] aims to offer an archival service for long-term preservation of herbarium specimen images and to develop innovative processes for extracting metadata from those images. HERBADROP follows the global trend towards scalable industrial-style digitizing of herbaria specimens. It is designed as both an archival service for long-term preservation of herbarium specimen images and a tool for analysing and extracting information written on the image, both supported by CINES [6], by using Optical Character Recognition (OCR) analysis.

Making the specimen images and data available online from different institutes allows cross domain research and data analysis for botanists and researchers with diverse interests (e.g. ecology, social and cultural history, climate change).

Herbaria hold large numbers of collections: approximately 22 million herbarium specimens exist as botanical reference objects in Germany, 20 million in France and about 500 million worldwide. High resolution images of these specimens require substantial bandwidth and disk space. New methods of extracting information from the specimen labels have been developed using OCR but using this technology for biological specimens is particularly complex due to the presence of biological material in the image with the text, the non-standard vocabularies, and the variable and ancient fonts. Much of the information is only available using handwritten text recognition or botanical pattern recognition which are less mature technologies than OCR [13].

The proposed platform is expected to support or even substitute costly manual data input as much as possible. The platform will also curate and enrich metadata resulting from image analysis using optical character recognition (OCR) and pattern matching.

Results are exposed as platform independent Web services which can be effectively integrated into herbarium data management systems as well as metadata capture workflows. Since 2016, five European community partners[1] have been involved. Their contribution to the pilot represents a business model that can be potentially replicated by other institutes.

The EUDAT B2SAFE service is used in the first step of the ingestion process. Existing images of herbarium specimens along with the associated metadata are transmitted to the CINES repository using B2SAFE transfer service. The ingestion into B2SAFE is carried out in accordance with the centralized persistent identifiers (PID) management system used in EUDAT. It is envisaged that discovery and visualization of the data objects will be performed with the EUDAT B2FIND service.

The data workflow in HERBADROP is represented by Figure 1.

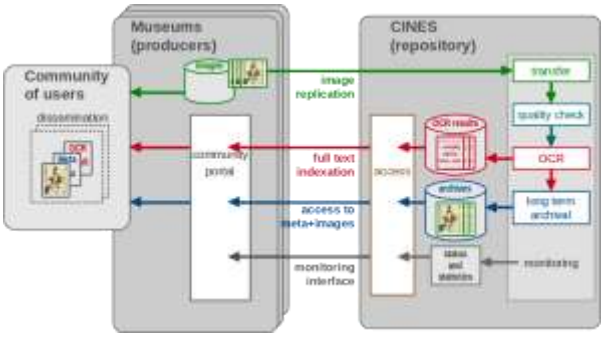


Figure 1 Data workflow of the HERBADROP data pilot

2.2 Data curation scenarios for HERBADROP

The HERBADROP communities have expressed their wish to implement specific use cases such as identifying duplicates amongst specimens from the different museums. This kind of requirement is very useful to improve EUDAT services. Another example of policy is long term preservation that involves a number of controls including file format verification and metadata quality. Amongst HERBADROP users, two partners of the community have proposed practical scenarios for data curation: Digitarium [14] and the Royal Botanic Garden of Edinburgh (RGBE).

Scenario proposed by Digitarium (Finland)

Digitarium [14] would like to use Optical Character Recognition (OCR) data to generate metadata based on the label information available for the herbarium specimen. Firstly, a Natural Language Processing based

[1]The partners in the HERBADROP data pilot are: Musée National d'Histoire Naturelle (MNHN) – Paris, France; Royal Botanic Garden of Edinburgh (RBGE) – United Kingdom; Botanic Garden and Botanical Museum (BGBM) – Berlin, Germany; Digitarium – Finland; Naturalis Biodiversity Center – Netherlands

system will be used to do OCR quality check and extract relevant terms. Then metadata will be either automatically generated, or manually inserted through the transcription portal [15] but with the help of OCR data.

More general for EUDAT infrastructure services, Digitarium would like to utilize and integrate them into the whole digitisation process of natural history biological collections. The data flow goes from the beginning of the digitisation process i.e. imaging, to storage, then to transcription and analysis, until accessing. This involves data storage, high-performance computing resources, and web services in EUDAT.

Firstly, the images from the imaging station can be transferred into EUDAT storage for long-term preservation instantly or in batch. After transferring, HPC can access the images and do OCR to extract label information to generate preliminary metadata. This metadata has to be associated with corresponding images. The data can be openly accessed. However, the access rights of data have to be set up for different purposes, such as endangered species protection.

Secondly, using HTTP APIs, the images and their metadata can be accessible from EUDAT by data-owner portals. Therefore, browsing and transcribing are available. Updated metadata will be transferred back into the EUDAT B2SAFE service. Different versions of metadata have to be kept.

Thirdly, the metadata is indexed. Therefore, the data can be searched or filtered based on different terms for further scientific usages. HPC resources can be utilized also on the data for different researches.

Scenario proposed by RBGE (the Royal Botanic Garden of Edinburgh) in association with MNHN (Musée National d'Histoire Naturelle) – Paris

The core of the concept of HERBADROP is to harvest metadata from OCR analysis of the text that is a part of herbarium images. The choice has been to proceed to a full text analysis using a Lucene-based engine Elasticsearch [16]. The objective of this approach is to provide a powerful interface for further data curation as part of the preservation process (identifying duplicates, or inducing new taxonomic relations, etc.), see [12].

Safeguarding long-term data storage is an important precondition for reliable access to herbarium specimen information. Thanks to this pilot, it is possible to envisage long-term storage for herbarium specimen images. Moreover, the specimens will be discoverable by the entire scientific community. Thus, undescribed species stored in herbaria can be examined by experts to aid identification and discovery of new species.

Distribution information for species over time can be evaluated and these data could provide evidence of the point in time when an invasive species first occurred in a certain area. Historians could analyse herbarium data to create itineraries for historical characters. The data can be used to calibrate predictive models of the oncoming changes in biodiversity patterns under global threats. This diverse information will be useful for a wide user community including conservationists, policy makers, and politicians.

3 GEOFON use case

The second use-case concerns GFZ, the German Research Centre for Geosciences. GFZ provides valuable seismological services in the form of a seismological infrastructure named GEOFON [7] to research and better understand our complex system Earth.

GFZ is one of the members of the EPOS initiative (European Plate Observatory System) [5] and, in this context, collaborates with other two seismological data centres related to EPOS (KNMI, INGV) in the EUDAT project.

Besides being one of the fastest earthquake information provider worldwide, GEOFON is also one of the largest nodes of the European Integrated Data Archive (EIDA) for seismological data under the ORFEUS[2] umbrella, which is a distributed data centre established to (a) securely archive seismic waveform data and related metadata, gathered by European research infrastructures, and (b) provide transparent access to the archives by the geosciences research communities.

The internal structure of GEOFON is based on three pillars:

- A global seismic network operated in close collaboration with many partner institutions with focus on EuroMed and Indian Ocean regions. The network consists of ca. 110 high quality stations, which acquire data in real time [8].
- A global earthquake monitoring system which uses data from GEOFON and partner networks [9]. It publishes most timely earthquake information. First automatic solutions are available few minutes after the events and mostly manually revised later.
- A comprehensive seismological data archive for GFZ and partner networks, for permanent networks as well as for temporary deployments.

For some GEOFON partner networks, GEOFON acts as a data centre saving a replica of the original copy and at the same time as a data distribution centre. Additionally, data from many temporary station deployments are permanently archived at GEOFON, in particular passive seismological experiments of the GFZ Geophysical Instrument Pool Potsdam (GIPP) and the German Task Force Earthquake.

Most data are open for public access, as well as real-time data feeds when available. However, there is a small amount of data under an embargo period, usually for a limited amount of time (3–4 years).

[2] Observatories and Research Facilities for European Seismology (http://www.orfeus-eu.org/)

3.1 Data workflow in GEOFON

GEOFON supports two scenarios for the ingestion of data into its archive: one for permanent networks and one for temporary (and most probably already finished) experiments.

Usually, raw data is transmitted to the data centre with the metadata (technical hardware description) to be able to operate with it. In the case of permanent networks raw data is received continuously from the stations around the world via satellite using a protocol called SeedLink [17], a real-time data acquisition protocol which works on TCP. The packets of each individual station are always transferred in timely (FIFO) order.

In the case of temporary experiments network operators provide usually, first, the metadata needed to use the data, and in a second phase the data to be archived. Data transmission can be done as in the permanent networks case (SeedLink protocol), or can also be transmitted to the data centre by the network operator using some client-server tools provided by GEOFON, which will do the first quality check of the data format. In some cases, both methods could be used.

A schematic view of the workflow at GEOFON can be seen in Fig. 2. It should be noted that this workflow is also valid for many of the seismological data centres belonging to EIDA/ORFEUS. For instance, the other two data centres piloting EUDAT services (KNMI and INGV).

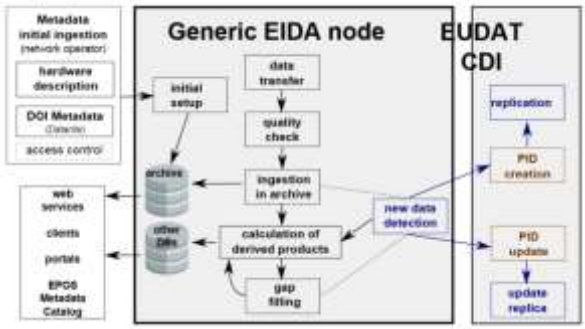


Figure 2 Data workflow from GEOFON. It also represents the workflows from a generic seismological data centre as the ones under the EIDA/ORFEUS initiative. Boxes in black are generic activities from the data centre. Blue boxes show activities related to the EUDAT service B2SAFE, while brown boxes show the tasks related to B2HANDLE

In both cases, permanent and temporary networks, data go through some quality checks after being received. When data are sent in real-time there is a first control by sorting the records before actually ingesting them into the archive (~1 day after reception). After 4–6 weeks, for stations that still have the buffered data, a gap filling process is started.

When data have been bulk uploaded to the data centre by the network operator, it is immediately checked to exclude overlaps. In this case, as all available data is copied off-line, there is no need to check for problems related to real-time transmission, like gaps and proper order of records, as they are checked by the automatic archiving tools.

In the case that the data is under an embargo period, the access control list is created or updated. After completion of the last steps, data is opened through standard access protocols.

The internal organization of the archive is based on an approach called SeisComP3 Data Structure (SDS). This means that files are stored under a predefined directory structure, which uses the codes from the network/station/channel used to record the data as well as the year. The continuous time series are stored in a standard seismological format called Mini-SEED. The time series are split in daily files for each recording sensor and, therefore, files are closed when the day finishes. At that moment, "new" data (recently closed files) can be processed to obtain derived products from them. For instance, quality metrics on the data or detailed availability information, which are offered to our users by means of a Web service.

Once the data is archived users can make use of any of the services provided by GEOFON to retrieve it. Considering that there are different services which can provide the data to the users, the usage statistics is centralized in one database to be able to analyse the impact of the data on the community regardless of the method used to retrieve it.

3.2 Service hosting environment with the inclusion of EUDAT services

Considering the workflow depicted in the previous section, GEOFON introduced some EUDAT services in order to automate and/or improve some of the tasks related to it.

Many services are being provided at GEOFON (e.g. interactive web portals, proprietary protocols to get data or derived products), with two of them (Station-WS and Dataselect) being particularly important, as they are international standards and the core services for the community upon which other services are built. Station-WS serves the information describing the hardware and everything related to the deployment, while Dataselect serves the data.

Two main EUDAT services have been integrated in the GEOFON workflow; namely, B2SAFE and B2HANDLE. The former is used to accomplish most of the Data Management tasks, while the latter is used to manage/store Persistent Identifiers (PIDs).

As the archive is stored in a directory structure from a partition, the B2SAFE service "mounts" the archive as an external resource in read-only mode.

One of the main requirements for the Data Policies at GEOFON is the capability to trigger processes based on the inclusion of new data. In the context of B2SAFE, this can be done by means of automatic rules which are executed under certain conditions (e.g. new data ingested).

With the proper rules we can enforce that, after new data is detected by B2SAFE, a certain set of actions is executed. For instance, the derived products can be generated and data can be replicated to a partner data centre from the EUDAT CDI, the Karlsruhe Institute of Technology (KIT). Also, as part of this replication

process, persistent identifiers (PIDs) are generated for each file, so that the PID can be used to globally and univocally identify the file.

PIDs are managed and stored by means of the already mentioned service called B2HANDLE, which is based on a Handle Server and other libraries developed within the project. GFZ has a broad expertise in this type of tools and, therefore, we decided to deploy our own B2HANDLE server and work with our local instance.

Each generated PID is stored with a set of key-value pairs called "PID Record". The information in the PID Record allows, among other things, to track other copies of the file in different data centres or validate its integrity by means of pre-calculated checksums.

3.3 Data Policies to apply at GEOFON through EUDAT services

After the formalization of the internal workflows at GEOFON, and the inclusion of requirements from the community and the data centre, we defined a set of Data Policies to be enforced by means of the tools available within EUDAT and new developments, which could be useful for different communities.

Some of them are related to the Replication process. For instance:

- replicate every new file in the archive to our internal backup server;
- if we are the official provider of the data in a file, replicate it to an off-site partner within the EUDAT CDI;
- seismological data that does not belong to us but comes from our earthquake early monitoring system should be kept for 6 months only; data still need to be replicated to the internal server;
- file deletion must not be possible in an automated way. In case that the system detects that a file should be deleted, an email should be sent to the appropriate operator.

Regarding the access control of the files:

- "Restricted data" must be tagged and proper access control must be applied to them;
- access restrictions can be automatically removed after a period of time (embargo period);
- data must be able to be accessed via an HTTP API respecting the ACL (Access Control List);

Regarding automatic metadata extraction:

- Metrics derived from the data must be automatically calculated to populate some of our services when new data is ingested.
- Detailed statistics related to the data access should be available for the data owners/creators.
- In case that data are modified (e.g. correcting errors, filling gaps), this information should be available for future use (provenance information).

Regarding the integrity of the stored data:

- a weekly process will select ~2% of the folders in our archive and verify that the synchronization is correct; the idea is that every file will be checked at least once in a year;
- check that the data is stored in SDS format;
- start and end time of network/station operation must be available and data outside this time span must not be allowed.

The identified relevant policies are being gradually implemented using generic EUDAT services and GEOFON-specific software.

4 Mapping of EUDAT data policies to SCAPE and RDA policy curation frameworks

For the design and implementation of data curation actions in EUDAT, the relevant outputs of SCAPE project [2] and Practical Policy Working Group of the Research Data Alliance [3] have been identified. SCAPE outputs are perceived of high quality owing to the advanced thinking that considered long-term digital preservation policies at a granular level suitable for the machine-executable implementation. RDA Practical Policy Working Group outputs are a result of a substantial international collaborative effort including experts in iRODS platform [11] that is a technological foundation of the EUDAT B2SAFE service.

For SCAPE, we used the catalogue of preservation policy elements [4] that is a systematized compendium of granular policies with examples of what SCAPE called "control policies" (granular statements that are easily translatable to machine-executable functions), and for the RDA Practical Policy Working Group it was their practical policy implementations report [9] that compiled a set of machine-executable functions for iRODS platform [11].

In addition to this top-down retrospective review of the SCAPE and RDA outputs, a bottom-up analysis of control policies applicable to the GEOFON and HERBADROP use case was performed, with a number of control policies identified as prime candidates for implementation in EUDAT B2SAFE. These policies are presented in Table 1.

Then the gap analysis was performed against SCAPE policy elements, to see whether these bottom-up identified control policies allow enough coverage of the extensively defined data curation policy landscape of SCAPE project. SCAPE policy elements catalogue [4] is two-level with Guidance Policies on the top level and Policy Elements on the granular level. An example of Guidance Policy is Authenticity Policy that breaks down to Integrity, Reliability and Provenance as policy elements. Hence control policies in Data Integrity checks category from Table 1 correspond to Integrity policy element of Authenticity Policy in the SCAPE policy elements catalogue.

One noticeable gap discovered through this mapping exercise is the Digital Object lifecycle which was paid due attention to in SCAPE policy landscape but is missing in the current EUDAT considerations. This gap may be hard to address as EUDAT is a collaborative project that accumulates data from a large variety of research communities with a wide range of digital object types and lifecycles. However, this discovery should

inform the future operation of EUDAT services so that they could meet all reasonable (and multi-aspect) requirements for data curation and long-term digital preservation.

Table 1 Candidate control policies for implementation by GEOFON and HERBADROP

Policy category	Control policy	Policy examples
Data replication	Number and location of replicas	Data should be replicated in N locations, including in locations A and B
	Timeframe for replication	Data should be replicated within the next 24 hours after the data ingestion in any particular location
	Data nodes roles	All data nodes are equivalent to read data from, but data can only be initially ingested in node X then replicated over all other nodes
Data integrity checks	The set of checksum algorithms acceptable	Checksum algorithm accepted is MD5
	Periodicity and scope of integrity checks	Calculate checksums for 2% of all data assets every week, with the aim of having the entire data collection checked annually
Data and metadata formats	Data formats accepted	BMP and PNG accepted for images
	Metadata extraction from data	Upon ingestion, file name should be extracted as metadata
	Data format check procedures acceptable	Software package X should be used for data format validation
	Minimal metadata assigned upon data release	PID is a mandatory metadata element
Data access and data reuse	Embargo rules	Embargo period of N years is applied to all PDFs and images
	The set of data licenses recommended upon data release	CC-BY license should be assigned to all data released after the embargo period ends
	Data reuse statistics collection	Number of file downloads should be collected

5 Conclusion and further work

Analysis of data curation requirements of two use cases: HERBADROP and GEOFON has been performed, coupled with the retrospective review of the elaborated data curation policies from a dedicated EU project (SCAPE) and practical (machine-executable) policies that were the output of the dedicated RDA working group.

A set of granular control policies have been identified as candidates for implementation in two use cases, and a gap analysis of these policies has been performed against the SCAPE catalogue of policy elements. A similar gap analysis should be performed against the RDA practical policies catalogue, in order to see what existing iRODS implementations can be reused for the creation of machine-executable policies in EUDAT B2SAFE service.

After the set of identified policies is applied in the two use cases that have been involved in their formulation, the same policy framework should be applied in a larger number of research communities associated with EUDAT through its pilot programme.

The scope of projects and initiatives in data curation and long-term digital preservation can be extended beyond SCAPE and RDA working groups; this specifically applies to popular functional models of digital preservation like OAIS [18] that we feel have not been thoroughly evaluated so far for their potential application in EUDAT.

The major result of these works is going to be a conceptually and terminologically consistent catalogue of machine-executable policies for EUDAT services that will be explicitly mapped to requirements of the participating research communities, as well as to mature data policy frameworks developed by EU projects and international collaborations dedicated to data curation and long-term digital preservation.

The EUDAT data policies catalogue will serve then both as guidance for machine-executable policy implementations and as a validation tool to ensure the compliance of EUDAT CDI services to high-level policies of data curation and long-term digital preservation. This should allow to promote certain EUDAT platforms such as B2SAFE from their current status of "bit-level" data management solutions topolicy-driven services where the actual set of policies can be configured according to a particular use case.

Acknowledgements

This work is supported by EUDAT 2020 project that receives funding from the European Union's Horizon 2020 research and innovation programme under the grant agreement No. 654065. The views expressed are those of authors and not necessarily of the project.

References

[1] EUDAT Collaborative Data Infrastructure. https://www.eudat.eu/eudat-cdi

[2] SCAPE: Scalable Preservation Environments. http://scape-project.eu/

[3] Research Data Alliance Practical Policy Working Group. https://www.rd-alliance.org/groups/practical-policy-wg.html

[4] SCAPE Catalogue of Preservation Policy Elements. http://scape-project.eu/wp-content/uploads/2014/02/SCAPE_D13.2_KB_V1.0.pdf

[5] EPOS: European Plates Observing System. https://www.epos-ip.org/

[6] CINES: French National IT Center for Higher Education and Research. https://www.cines.fr/en/

[7] Hanka, W., Kind, R.: The GEOFON Program. Annals of Geophysics, 37 (5), Nov. 1994. ISSN 2037-416X. doi:10.4401/ag-4196

[8] GEOFON Data Centre (1993): GEOFON Seismic Network. Deutsches GeoForschungsZentrum GFZ. Other/Seismic Network. doi: 10.14470/TR560404

[9] Practical Policy Implementations Report. http://dx.doi.org/10.15497/83E1B3F9-7E17-484A-A466-B3E5775121CC

[10] Hanka, W., Saul, J., Weber, B., Becker, J., Harjadi, P., Fauzi and GITEWS Seismology Group: Real-time Earthquake Monitoring for Tsunami Warning in the Indian Ocean and Beyond, Nat. Hazards Earth Syst. Sci., 10, pp. 2611-2622 (2010). doi:10.5194/nhess-10-2611-2010

[11] iRODS: Integrated Rule-Oriented Data System. https://irods.org/

[12] Haston, E., Chagnoux, S., Dugénie, P.: Herbadrop – Long-term Preservation of Herbarium Specimen Images. Proc. of the second Eudat User Forum. Rome (2016). https://www.eudat.eu/communities/long-term-preservation-of-herbarium-specimen-images

[13] Dugénie, P., Chagnoux, S.: EUDAT Data Pilot Herbadrop. Second Interim Herbadrop Data Pilot report (2016)

[14] Digitarium: Service Centre for High Performance digitization. http://digitarium.fi/en

[15] DigiWeb+digitization platform. http://digiweb.digitarium.fi/

[16] Elasticsearch Search and Analytics Engine. https://www.elastic.co

[17] SeedLink Protocol and Tools Overview. http://ds.iris.edu/ds/nodes/dmc/services/seedlink/

[18] Reference Model for an Open Archival Information System (OAIS), Recommended Practice, CCSDS 650.0-M-2 (Magenta Book). Issue 2, June 2012. CCSDS (The Consultative Committee for Space Data Systems), Washington DC (2012). EUDAT services. https://www.eudat. eu/services-support

[19] EUDAT services. https://www.eudat.eu/services-support

Data policy as activity network

© Vasily Bunakov

Science and Technology Facilities Council, Harwell Campus,
United Kingdom

vasily.bunakov@stfc.ac.uk

Abstract. The work suggests using a network of semantically clear interconnected activities for a formal yet flexible definition of policies in data archives and data infrastructures. The work is inspired by needs of EUDAT Collaborative Data Infrastructure and the case of long-term digital preservation but the suggested policy modelling technique is universal and can be considered for all sorts of data management that require clearly defined policies linked to machine-executable policy implementations.

Keywords: data management, long-term digital preservation, data policy, semantic modelling.

1 Introduction

Problematics of advanced long-term digital preservation [1] has been in focus of many collaborative projects and popular recommendations. However, it has been paid a relatively small attention in domain-specific projects that rely on data archiving, or in projects that develop scalable e-infrastructures aggregating data that comes from different user communities.

One of the problems that long-term digital preservation aims to address is having clear policies for the entire data lifecycle from data ingestion by archive or by e-infrastructure service, through years-long data management with sensible data checks, transformations and moves, to data access and data dissemination to the end users.

One can argue that without clear data policies and means of their validation there is no such a thing as the long-term digital preservation, even in cases when a technology foundation used for an archive or an e-infrastructure is sound and well-supported. At the end of the day, every technology evolves – and at a brisk pace compared to relatively long time when many data assets are going to be useful, so data policies and means of their expression should be semantically clear and in a way more permanent than technology that underpins data management. A strong case for policy-driven digital preservation, with extensive references to the prominent projects and popular methodologies was made in [2].

In practice, quite a few data archives and e-infrastructures end up in a situation when they have got a sound technology for managing data bits, also acquire a decent number of users (which is a popular measure used by funders for their judgement on the e-infrastructure success) but do not have a reasonable data policy, let alone any machine-assisted reasoning over it. The users' trust in the archive or the e-infrastructure may be enough for their daily use but there can be a substantial conceptual and technological gap in regards to data policies formulation, expression and execution.

Some larger projects and e-infrastructures are aware of this gap and do make efforts to close it by working on data policies implementation. An example of such e-infrastructure is EUDAT [3] that has developed a number of operational services [4] and data pilots with user communities, and is now trying to express and apply policies to these services.

The prime candidate for applying data policies in EUDAT is B2SAFE service [5] based on iRODS platform [6]. B2SAFE developers are doing a very good job on building geographically and organizationally distributed data storage with data replication, integrity checks and other routine tasks of data management guided by iRODS machine-executable rules. B2SAFE have made their own effort on policies with the development of Data Policy Manager [7] which is a software module with policies expressed via XML templates. There is a perceived need though of having a more universal solution for policy management across all EUDAT services. The possible policy modelling approaches under consideration are using RuleML[8], SWRL[9] or ProvOne ontology[10] which seems suitable not only for capturing data provenance after the execution of certain actions but also for the forward-looking design of data processing workflows which can then potentially serve as a means of data policy modelling.

This work presents an alternative approach to those mentioned and is based on Research Activity Model [11] which is in fact quite universal and suitable for the expression of all sorts of activities, not necessarily related to research. Research Activity Model is slightly extended and applied to the case of data policy modelling.

The main advantage of this alternative approach is its high modularity which allows modeling policy elements and using them as building blocks for the semantically clear representation of a whole policy. The modularity of policy design is especially important in data infrastructures that commonly aggregate data coming from different user communities, often having their own business models, technical requirements, data formats and data lifecycles which makes it difficult to design and adequately express the crosswalks between community-specific data policies and those for the data infrastructure. Another advantage of the suggested approach is its ability to address the conceptual gap between policy formulation and policy implementation, as it may not be easy to

Proceedings of the XIX International Conference "Data Analytics and Management in Data Intensive Domains" (DAMDID/RCDL'2017), Moscow, Russia, October 10–13, 2017

translate a high-level policy (often in a textual form) into machine-executable policy.

The modularity should allow high levels of inheritance and reuse of policy elements; it also helps to solve specific problems of policy formulation and validation when textually the same policy can be executed in different ways leading to different states of data archive, for which situation we provide an example. The conceptual gap between policy formulation and policy implementation is addressed by a possibility to define policy-related Activities as "black boxes" with (initially) only interfaces defined; this can be hopefully done by policy makers themselves without entirely delegating this policy design phase to policy implementers (software developers).

Implementation of a sensible data policy is a challenging task even within the boundaries of a particular organization. In a situation when the organization is using a collaborative data infrastructure along with its own organization-specific IT services, the implementation of a data policy is going to be even more intricate and is likely to rely on loosely coupled services. An approach to data policy modelling suggested in this work is going to address this challenge, along with the alleviation of the earlier mentioned problems of the policy elements reusability and the policy application results predictability.

The work is inspired by needs of EUDAT Collaborative Data Infrastructure [3] and refers to it for illustration of certain ideas, also the main incentive for the work was modelling policies for the case of long-term digital preservation. However, the suggested modelling technique is universal and can be considered for all archives or e-infrastructures that are interested in all sorts of data management (not only long-term digital preservation) that require a clearly defined policy linked to machine-executable policy implementations.

Conceptual challenges of data policy modelling are discussed first, specifically the problem of policy decomposition into policy elements, then an example is given of how Activity Model can be used for policy modelling. This is followed by suggestions on what IT architecture for data policy management will be required to support the suggested modelling techniques.

2 Data policy and a problem of its decomposition

2.1 Insufficiency of granular policy definition

Data policy is often created as a conventional textual document that contains certain statements about what should or should not be done with data, with implied or sometimes explicit logical "ANDs" and "ORs" that glue statements together in an aggregated policy. This composite nature of policies is why it seems natural to break down the policy document into granular statements, model each statement using some formalism and then execute the statements using some IT solution.

One of the most advanced efforts on data policy decomposition was performed by SCAPE project [12] that created an extensive catalogue of preservation policy elements [13] which are in fact granular textual statements. These granular statements which can be converted, in a pretty straightforward way, in machine-executable statements are called *control policies* in SCAPE. Examples of control policies are: "information on preservation events should use the PREMIS metadata schema" or "original object creation date must be captured". The granular control policies relate to a higher-level *procedural policy* (a procedural policy on Provenance for the current example) which in turn relates to an even higher-level and most abstract *guidance policy* (a policy on Authenticity for the current example). Three-level structure of guidance policies, procedural policies and control policies constitute a very well developed SCAPE digital preservation policy framework.

SCAPE stopped short of the actual implementation of control policies, so when EUDAT [3] decided to use the SCAPE framework for policy considerations, it was also decided to supplement this framework with the catalogue of practical data policies [14] developed by an RDA (Research Data Alliance) Practical Policy Working Group. The practical data policies in this catalogue are expressed as iRODS [6] functons specifically suitable for implementation in EUDAT B2SAFE service [5] based on iRODS platform.

Having well-defined control policies or practical policies is not enough though for semantically clear modelling of a data policy as a whole, as the application (execution) of a policy composed of granular machine-executable statements may lead to quite different outcomes depending on the order in which granular policies are applied.

The problem of policy decomposition is in fact interrelated with the problem of policy validation. To illustrate this, let us consider a simple case when there is a couple of easily identifiable policy statements contained in the same policy document which we want to decompose and validate through execution of two granular policies. Let the statements in a composite policy (perhaps, but not necessarily so, added one to another through some policy update by different policy managers) be:

[1] Image files having size of more than X gigabytes should be stored in file storage A; otherwise they should be stored in file storage B.

[2] Image files of type RAW should be converted in JPG format.

If a certain file of type RAW is more than X gigabytes in size but becomes less than X when converted in JPG then, depending on the higher-level guiding policy and on the order in which these granular policies are applied in the actual service implementation, the result of the combined application of the two granular policies can be any of the following:

1. File is moved as RAW in storage A and remains stored in A as RAW.
2. File is moved as RAW in storage A then converted in JPG and remains stored in A.
3. File is converted in JPG and stored in B.
4. File is moved as RAW in storage A and remains stored in A as RAW; also a copy of it converted in JPG is stored in B.

This is to illustrate that validation of the data policy

implementation is hard as any of the listed outcomes may be considered being right or wrong depending on the validator's point of view.

Also let us take into account that policy validation can be based on some statistical selection of samples (so that problematic boundary cases of RAW data sized only slightly over X gigabytes threshold may not be selected in a sample and hence go unnoticed), or that a policy validation procedure allows some tolerance towards small amount of failed policy checks (so that even if a few files have ended up somewhere that a particular policy interpretation considers to be a wrong place, this does not trigger a policy violation alert).

So even if the data policy can be, seemingly successfully, decomposed into granular policies that are easy to define and validate as machine-executable statements, the actual result of the policy implementation does not necessarily match the intentions of policy designers or policy managers, as the backwards process of the policy composition – assembling it from the granular policies (policy elements) – can be performed with substantial variations.

2.2 Possible responses to the challenge of granular policies insufficiency

One possible response to the outlined challenge could be setting up an elaborated policy governance framework, i.e. well-defined business processes that allow human agents (policy managers) to look after the policy implementation, i.e. accumulate and analyse feedback from the environment where the policy is applied and supply the result of this analysis as updated requirements to software developers who work on the actual software implementation of the policy. This approach requires a good organizational culture and a substantial human resource involved in data policy management and in policy implementation; documented requirements will serve as an interface between policy managers and policy implementers. Some "magic" should happen in between so that high-level policy definitions translate into actual policies implementation in software code, this is why policy validation is likely to demand extensive software testing with specific policy-related test cases.

Another possible response is having an elaborated means of expression for the entire data policy (a sophisticated policy modelling language): both for the definition of granular policies and for the definition of logic than binds the granular policies into the whole. An example of this approach is RuleML [8] that is considered a candidate for a detailed expression of data policy in EUDAT e-infrastructure [3]. This approach requires skilled human resource for policy modelling; the modeler and a sophisticated model produced by her becomes then an interface between policy managers and policy implementers (the role of the latter is less prominent than in the first approach, in a sense that software developers should not interpret requirements but just implement – or adopt – a certain engine that executes formal rules defined by the savvy policy modeller).

The third possible response is that a certain formalism is used for the expression and, where necessary, recomposition of granular policies (policy elements) and for their assembling in the whole, with that formalism being reasonably friendly to machines as well as to humans. The humans – policy managers themselves or a not-so-skilled modeller – can use the formalism for a flexible policy definition that can be fairly easily modified depending on the true policy intentions and on the feedback received from the archive or e-infrastructure where the policy is implemented. The role of software developers is then to implement an engine for the formalism (quite similarly to the second approach). The machine just executes the policy expressed using that formalism.

The differences amongst approaches are presented in Table 1; in essence, they are different "weights" (different levels of demand) for the skills of policy managers, policy modellers and policy implementers.

Table 1 Differences amongst policy modelling approaches

Policy modelling approach	**Demands for policy manager skills**	**Demands for policy modeller skills**	**Demands for policy implemen-ter skills**
Policy governance framework + requirements management + specific software testing	High	None (policy modeler can be replaced by business analyst or/and software tester)	High
Policy modelling language	Low	High	Medium
Formalism for granular policy elements definition and composition	Medium	Medium	Medium

The preferable approach could easily be the third one as it empowers policy modelers themselves with reasonable means of policy expression and therefore can reduce overheads and risks of communicating a policy from policy managers through modelers to implementers. A remote analogy of the third approach could be the proliferation of SQL language that, despite its sophistication, has become a lingua franca of not only software engineers but is widely used by logistics and even sales departments is all sorts of business.

The formalism to be used for data policy expression

should not be something as developed as SQL though, neither should it be purely textual: it can be based on the idea of "building blocks" with possible graphical representation of them, hence providing an easy-to-operate semantic wrapper for machine-executable statements. On the other hand (unlike SQL which allows the actual data manipulation), these "building blocks" for data policy definition are likely to remain only a wrapper to the actual machine-executable implementations of granular policies which will be inevitably specific to a particular service even within the same archive or e-infrastructure. As an example, for EUDAT B2SAFE [5] that is based on iRODS platform [6] these granular implementations can be iRODS functions and for other EUDAT services based on other software platforms the policy implementations can be something else. A common semantic wrapper will be then a reasonable means of a clear policy modelling and a clear definition of interfaces between policy "building blocks" across a variety of different IT services.

This work strongly prefers the third approach and suggests considering Activity Model [11] for semantically clear modelling of data policies in all IT services within the same data archive or e-infrastructure, as well as for policy interoperability across different data archives and e-infrastructures.

3 Activity Model as a semantic wrapper for machine-executable policies

3.1 Activity Model in a nutshell

Activity Model [11] was initially suggested for modelling granular research activities and combining them in networks so that, as an example, the output of one Activity can be the input of another one, e.g. these combined Activities may represent certain phases in research data analysis. It has been clear though that Activity Model can suit all sorts of activities as it is pretty generic; as an example, it may well suit for modelling data provenance across different IT services within e-infrastructure.

The main "building block" of the Activity Model is an "activity cell" represented by Figure 1 with its aspects (that can be thought of as incoming and outcoming relations) explained in Table 2.

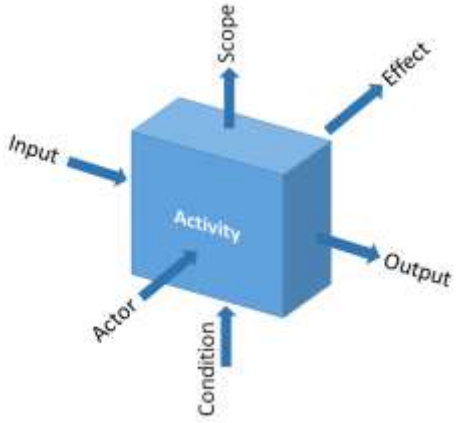


Figure 1 Research activity "cell"; it can be used for semantic definition of any activity

The full RDF serialization of the Activity Model is published in [11]; it is really simple and requires only RDF Schema and an "inverseOf" OWL statement for its expression, i.e. what is often referred to as RDFS Plus.

Table 2 Activity Model aspects explained

Aspect	Description	Examples	
		Research per se	Research data analysis
Input	Something that is taken in or operated on by Activity	Previous research	Raw data
Output	Something that is intentionally produced by Activity	Raw data	Derived (analyzed) data
Scope	Something that Activity is aimed at or deals with	Sample properties	One or more experiments
Condition	Something that affects or supports Activity, or gives it a specific context	Scientific instrument	IT environment
Actor	Something or somebody who participates in Activity	Investigator	Data analyst
Effect	Something that is a consequence of Activity	Environment pollution	New software module

Activity "cells" can be combined in chains or networks, and not necessarily in a way that the Output of one Activity is the Input to another. As an example, a data management policy can be the Output of one Activity (policy design) and the Condition that affects another Activity, e.g. data replication in the archive.

The model flexibility when any aspect of one Activity can be matched with any aspect of another Activity is supported by the fact that aspects do not have to have types associated with them.

3.2 Proposed extensions of the Activity Model

In order to use Activity Model for data policy modelling, we will need to make a profile of the model by specifying certain types of Activity as subclasses (in

case of an RDF serialization of the model – RDFS subclasses). Suggested extensions are presented in Table 3. Conceptually, Generic Data Management Activities should cover the needs of data engineering that are related to machine-interpretable policy implementations, Logical Switch Activities should cover the needs of data analysis and machine-assisted reasoning, and Control Activities should cover the needs of IT services deployment and operation.

Compared to modelling data policies with workflows, the suggested approach based on the definition of policy-related Activities should allow more loosely coupled implementations of policy management IT solutions. As an example, the "data engineering" part of policy implementation represented by Generic Data Management Activity can be performed on a software platform fully controlled by a specific user community or organization (e.g. a research institution), the operation (the actual execution of control statements) represented by Control Activity can be performed by collaborative data infrastructure (e.g. by EUDAT CDI [3]) and the logic of combining policy elements represented by Logical Switch Activity can be performed by either the organization or the data infrastructure, or by a third-party service.

If the policy was modelled by an executable workflow, it would require the presence of all three aspects: data engineering, reasoning and execution – in the same workflow likely operated by a single universal workflow engine. This would mean not only an operational limitation but a conceptual / modelling limitation, too, as all the participants (stakeholders) of policy implementation would have to adhere to the conceptual framework and the format required by the workflow engine. Modeling with interconnected Activities as semantic wrappers to particular implementations leaves more freedom to conceptualize and to operate data policies that are going to be executed by loosely coupled IT services.

Table 3 Additions to the core Activity Model required for data policy modelling

Type to add	Comment / Description
Generic Data Management Activity	Subclass of Activity for data policy definition. It can be considered a semantic wrapper for a variety of data handling Activities, e.g. Activities for data characterization or data transformation.
Logical Switch Activity	Subclass of Activity for logical switches of all sorts
Control Activity	Subclass of Activity for an interface with a particular software platform where policies are executed. This is a semantic wrapper for the actual call to a platform-specific script or function.

Depending on a particular operational environment (software platform where policies are executed), other parts of the Activity Model, e.g. its Inputs, Outputs, or Conditions may require additional semantically clear extensions. However, it is unclear at the moment whether these potentially required extensions should be a part of the universal Activity Model profile for data policies, or it is better to introduce them as necessary, as parts of policy execution engine implementations on particular software platforms.

3.3 Examples of the Activity Model data policies profile application

The role of the suggested model extensions will be clearer by giving an example of their application to the modelling of a particular policy. The example will be a policy with two granular statements about data movements depending on data size and data format that were considered in Section 2.1.

We will need to define first a File Characterization Activity:

```
   @prefix am:
   <http://.../stuff/ActivityModel#> .
   @prefix ampp:
   <http://.../ActivityModel#PolicyProfile> .
   GDMA_FileChar                                a
ampp:GenericDataPolicyActivity
GDMA_FileChar am:hasInput File
GDMA_FileChar am:hasOutput FileSize
   GDMA_FileChar am:hasOutput FileFormat
   GDMA_FileChar am:hasOutput File
   GDMA_FileChar am:hasScope ImageFiles
   GDMA_FileChar                 am:hasCondition
ServiceInstance
   GDMA_FileChar am:hasActor CertainScript
   GDMA_FileChar am:hasEffect FileCharLog

   In short, GDPA_FileChar activity takes a file
as an input and produces values for the file
size and file format (which can be semantically
clearly defined as necessary – e.g. with
measurement units and format IDs in a file type
registry) as outputs; the initial file is passed
over as another output. To derive the file size
and format, the activity uses CertainScript
(which again can be semantically clearly defined
as necessary – e.g. with references to a software
repository).
   As an additional outcome (better defined not
as Output but as Effect) of the file
characterization activity, we get the
FileCharLog log file. The scope of activity is
defined as ImageFiles (so that other kinds of
files can be handled by differently defined
Characterization Activities; what "ImageFiles"
actually means can be clearly defined with e.g.
a reference to a certain taxonomy entry). The
Condition is defined as ServiceInstance (which
means that Actor:CertainScript operates in some
particular IT service environment).
   Mapping of Activity to a particular software
implementation can be performed using Activity
ID and a reference to a repository with a clear
software identity, e.g. a software versioning
repository.
```

The graphic representation of this Characterization Activity (which, in the ideal world, can be designed in a certain authoring tool with graphical user interface and producing the above RDF as a serialization) is illustrated by Figure 2.

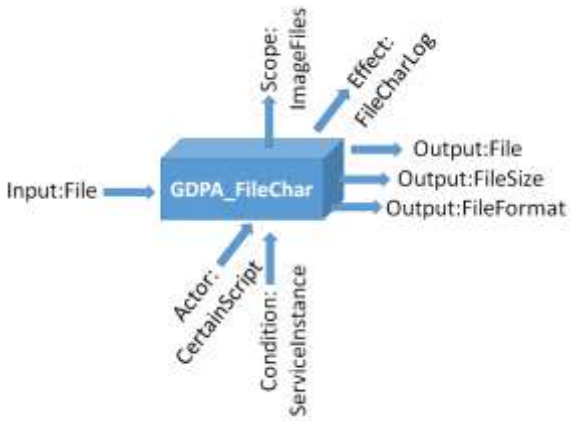


Figure 2 Definition of a Data Policy Activity for image files characterization

The problem of the policy composition out of two granular policies outlined in Section 2.1 can be addressed with the help of other classes of activities that we introduced earlier: Logical Switch and Control. For the sake of simplicity (as we are going just to illustrate it how the policy modelling can be done) we will not be defining all aspects for these activities, e.g. we can omit Scope or Effect but they may be required in a real policy modelling situation.

The Logical Switch activity will take File, FileSize and FileFormat as Inputs, a particular logic of handling file moves to either storage A or B, as well as file conversion, will be Condition. The Activity yields a list of particular control statements (like "move File to storage A", "Convert file in JPG format") as Output. The shape of such defined Logical Switch activity is illustrated by Figure 3.

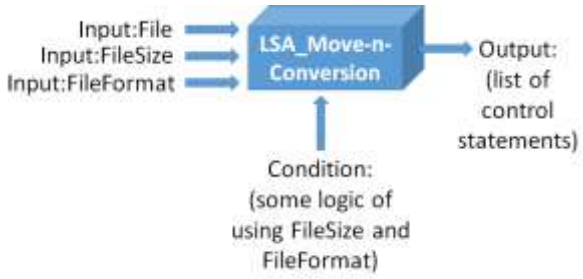


Figure 3 Definition of a Logical Switch Activity for handling image files

The semantically clear definition of a Logical Switch Activity gives an idea of how we suggest to address the problem of a policy composition from granular policy statements. The hope is, if the logic of producing control statements is made explicit, as well as the control statements themselves, this will eliminate the ambiguity of a policy composed of granular policy statements.

A good question is what formalism, if any, will be adequate for the expression of logic in the Condition of the Logical Switch. The short answer is: it depends on the policy engine implementation. In an extreme case, this Condition can be just a mandatory textual explanation (commentary) of the logic implemented by the Actor (which is omitted in the Figure 3), i.e. by an executable function or a procedure or a script for a particular IT platform. Alternatively, rules modelling language or workflow templates (and appropriate engines for them) can be used – yet, in this case, the actual usage of these modelling languages or workflow templates would be limited to the policy logic enwrapped in the Logical Switch Activity, allowing freedom for different implementations of other types of Activities involved in the policy definition.

How to express control statements in the Output is subject to particular implementations, too. The only consideration which is important for the moment – important both from conceptual and from implementation perspectives – is having the list of control statements as a clearly defined interface between Logical Switch Activity and Control Activity.

Control Activity takes the list of control statements as Input and makes platform-specific function or procedure or script calls that implement the control statements. Actors for Control Activity are particular functions / procedures / scripts and the Effects of it are log and error files or messages – whatever is used for traceability in a particular implementation. Condition is, similarly to the file characterization activity definition, a particular software platform or IT service where Actors operate. Figure 4 presents an example of a diagrame for the Control Policy.

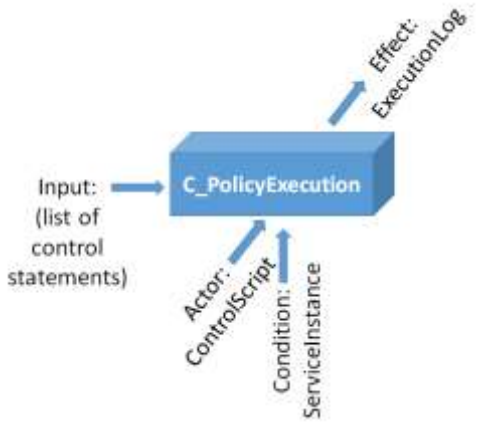


Figure 4 Definition of a Control Activity for policy execution

Generic Data PolicyActivities (such as data characterization) can be combined with Logical Switch Activities and Control Activities in a chain or a network of activities. For our example, the resulted chain is illustrated by Figure 5. It represents the full model of a certain data policy expressed as a chain of semantically clear activities with interfaces between them, as well as interfaces to activity implementations in particular IT services or software platforms.

It is worth mentioning once again that every aspect in the Figure 5 diagrame (such as File, Size, Format, Script or Log) should be thought of not as a particular artefact or a value but as a semantic wrapper of an artefact or a value. As a particular model serialization, these semantic wrappers can be RDF statements about artefacts or values.

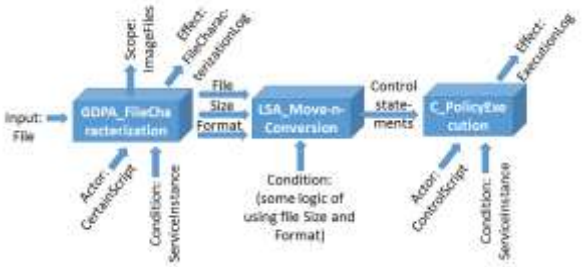


Figure 5 Example of full policy definition

In real data policy modelling situations, it may be necessary to define more than one instance of each Activity type; as an example, there could be two Data Characterization Activities defined (one for the file size and another for the file format) in place of one in our example. Nevertheless, even differently defined Activities could be combined in a semantically clear network representing the same data policy.

If Activities in Figure 5 are clearly defined and sensibly combined in the Activity network, this eliminates any ambiguity in policy definition and execution exemplified by two interfering granular policies discussed back in Section 2.1 so that the actual result of the policy implementation becomes predictable and can be formally validated.

One of the strengths of the suggested model is a combination of its reasonable expressivity with its high flexibility as it is based on the idea of composition of activities that can be a) modelled differently b) implemented differently and c) operated (executed) differently. In the above example, scripts for file characterization and scripts for policy execution can be implemented using different software and operated by different components of the same service, or by different services, or even by different e-infrastructures.

The actual chain or network of activities, as well as definition of each of them (i.e. definition of all semantic wrappers) could be done in a certain authoring tool with a graphic user interface and RDF as a model serialization format. Development of such a tool has been beyond resources available for this conceptual work; however, such a tool is worth mentioning as one of the elements of an IT architecture that can support data policies formulation, execution and validation.

4 IT architecture for activity-based data policy management

The proposed IT architecture is presented by Figure 6 with the most essential components and information flows (that would constitute a core operational platform for data policy management) designated as filled-in boxes and arrows; more advanced components and flows are designated as dashed boxes and arrows with a blank background.

As already suggested, having policy Activities authoring tools with GUI and possibility to serialize Activity networks in a semantically explicit format such as RDF is essential for good levels of adoption of the suggested approach and therefore such authoring tools should be a part of a sensible IT architecture for data policy management. In addition, what is required is a repository where policy designs can be stored and retrieved from.

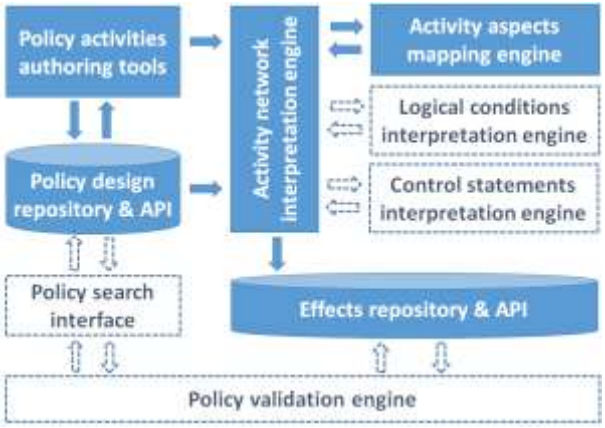


Figure 6 IT architecture for activity-based policy management

Activity network interpretation engine picks up Activity network from the authoring tools or repository and executes them. In order to execute activity networks in a particular IT environment (software platforms and services), a mapping engine is required that maps Activities and their aspects (such as Conditions or Outputs) to configuration files and executable scripts.

In addition to this generic mapping engine, specific engines for logical conditions and control statements can be implemented. Effects repository stores Effect aspects of each Activity; it is a generalization of logging service and contains semantically clear tracks of Activities execution. Policy search interface can be designed for searching and sharing data policies.

For the purposes of data archive or data infrastructure audit, a policy validation engine is required that talks to policy search interface and to Effects repository. The actual validation can be based on matching graphs of artefacts resulted from policies execution with graphs of Activities in the policy design.

5 Conclusion

The problem of data policy modelling with reasonable crosswalks between high-level (read: textual) policies and their machine-executable implementations has yet to find a satisfactory solution. The challenges of policy design and implementation are even bigger when collaborative data infrastructures are operated in combination with the in-house software platforms.

The problem of semantically clear crosswalks and the problem of data policy implementation across organization-specific and external IT services can be addressed by adoption of certain policy modelling techniques and tools. Activity Model [11] can be a reasonable means for the design of such tools, with the idea that data policies can be represented as networks of Activities with interconnected aspects of them.

This work has introduced extensions to the Activity Model in order to make it fit for the task of data policy

modelling. An example of using the Activity Model for the definition of a particular data policy has been given, and a possible IT architecture has been considered that can support data policy management based on Activity networks.

Acknowledgements

This work is supported by EUDAT 2020 project that receives funding from the European Union's Horizon 2020 research and innovation programme under the grant agreement No. 654065. The views expressed are those of the author and not necessarily of the project.

References

[1] Giaretta, D. Advanced Digital Preservation. Springer, Heidelberg (2011)

[2] Bunakov, V., Jones, C., Matthews, B., Wilson, M. Data authenticity and data value in policy-driven digital collections. OCLC Systems & Services: International digital library perspectives, vol. 30 issue 4, pp. 212-231 (2014). doi: 10.1108/OCLC-07-2013-0025. Open Access version of the preprint: http://purl.org/net/epubs/work/12299882

[3] EUDAT Collaborative Data Infrastructure. https://www.eudat.eu/eudat-cdi

[4] EUDAT services. https://www.eudat.eu/services-support

[5] EUDAT B2SAFE service. https://www.eudat.eu/b2safe

[6] iRODS: Integrated Rule-Oriented Data System. https://irods.org/

[7] EUDAT Data Policy Manager. https://github.com/EUDAT-B2SAFE/B2SAFE-DPM

[8] RuleML Wiki pages. http://wiki.ruleml.org/index.php/RuleML_Home

[9] SWRL: A Semantic Web Rule Engine. https://www.w3.org/Submission/SWRL/

[10] ProvONE: A PROV Extension Data Model for Scientific Workflow Provenance. http://vcvcomputing.com/provone/provone.html

[11] Bunakov, V. Core semantic model for generic research activity. In 15th All-Russian Conference "Digital Libraries: Advanced Methods and technologies, Digital Collections" (RCDL 2013), Yaroslavl, Russia, 14-17 Oct 2013, CEUR Workshop Proceedings (ISSN 1613-0073) 1108, 79-84 (2013). Persistent URL: http://purl.org/net/epubs/work/10938342

[12] SCAPE: Scalable Preservation Environments project. http://scape-project.eu/

[13] SCAPE Catalogue of Preservation Policy Elements. http://scape-project.eu/wp-content/uploads/2014/02/SCAPE_D13.2_KB_V1.0.pdf

[14] Practical Policy Implementations Report. http://dx.doi.org/10.15497/83E1B3F9-7E17-484A-A466-B3E5775121CC

Модель рекомендательной системы на нечетких множествах как эффективное расширение коллаборативной модели

© Д.М. Понизовкин

IT-Aces,
г. Переславль-Залесский, Россия

denis.ponizovkin@gmail.com

Аннотация. Рассмотрены рекомендательные системы, использующие коллаборативную фильтрацию для решения таких задач, как определение степени близости объекта пользователю (задача прогнозирования) и определение подмножества объектов мощности N, близких пользователю (задача $topN$). Такие системы считаются хорошо изученными и успешно применяются в коммерции, однако существуют открытые проблемы, связанные с использованием таких систем. Эти проблемы описаны в настоящей работе. В качестве метода устранения существующих недостатков предложена модель рекомендательных систем, которая основана на теории нечетких множеств и использует методы коллаборативной фильтрации.

Ключевые слова: рекомендательная система, коллаборативная фильтрация, мера сходства, отношение близости, эффективность, нечеткая логика.

The Model of Recommender Systems based on Fuzzy Logic as the Extension of the Collaborative Filtering Model

© Denis M. Ponizovkin

IT-Aces, Pereslavl-Zalessky, Russia

denis.ponizovkin@gmail.com

Abstract. In this article, we analyze collaborative filtering. We show existing problems connected with using of collaborative filtering. We propose the recommender system model based on the fuzzy logic theory. This model is the extension of the collaborative filtering which removes described problems.

Keywords: recommender system, collaborative filtering, similarity measure, fuzzy logic.

1 Терминология и обозначения

Рекомендательные системы (далее РС) [1] – одна из развивающихся областей Computer Science, начавшая свое существование с конца прошлого столетия [2]. РС являются инструментом, который облегчает пользователю задачу поиска нужной информации путем предоставления рекомендации по использованию соответствующей информации или за счет определения степени близости конкретной информации интересам пользователя.

РС работают со следующими исходными данными:

- $u \in U \subset \mathbb{N}$ – идентификаторы пользователей РС;
- $i \in I \subset \mathbb{N}$ – идентификаторы объектов предметной области РС, например, фильм в РС в области кинематографии; для простоты изложения не будем каждый раз употреблять выражения «пользователь» или «объект», будем обозначать их кратко «объекты»;
- $\rho: U \times I \to [0,1]$ – функция оценки близости и объектов; значение $\rho(u,i)$ показывает, насколько объекты i и u близки; как правило, оценки близости задаются самими пользователями за время работы с РС; будем считать, что чем меньше значение оценки, тем объекты ближе; будем говорить, что между пользователем u и объектом i выполняется отношение близости $\mathcal{R}$, если $\rho(u,i) \le \varepsilon_0 \in \varepsilon(0)$; будем называть такие объекты близкими.

Как правило, РС решают следующие две задачи (пользователь, для которого производится решение, называется *активным* и обозначается символом u_a):

Труды XIX Международной конференции «Аналитика и управление данными в областях с интенсивным использованием данных» (DAMDID/ RCDL'2017), Москва, Россия, 10–13 октября 2017 года

1. *Задача прогнозирования*: спрогнозировать неизвестное значение $\rho(u_a, i_p) = \perp$ (символом $\perp$ будем обозначать неизвестное значение) путем алгоритмического вычисления значения прогнозной функции $\overline{\rho}(u_a, i_p): U \times I \to [0,1]$, где i_p – прогнозируемый объект; при этом требуется, чтобы прогноз был составлен точно, то есть $|\overline{\rho}(u_a, i_p) - \rho(u_a, i_p)| \leq \varepsilon_0$;

2. Задача $topN$ – формирование подмножества объектов

$$I_{topN} = \{i: (u_a \mathcal{R} i) \wedge \rho(u_a, i) = \perp\} \wedge |I_{topN}| = N.$$

Так как неизвестно, выполняется ли отношение $u_a \mathcal{R} i$ в силу того, что $\rho(u_a, i) = \perp$, то выполнение отношения $u_a \mathcal{R} i$ определяется по значению прогнозной функции: $u_a \mathcal{R} i \Leftrightarrow \overline{\rho}(u_a, i) \leq \varepsilon_0$. Решение названных задач производится РС за счет анализа информации о характеристиках пользователей и объектов. X – множество характеристик пользователей, например, социально-демографические показатели банковской РС. Y – множество характеристик объектов, например, наименования кинематографических жанров. Значением характеристик пользователей является значение весовой функции $w_U: U \times X \to [0,1]$, объектов – $w_I: I \times Y \to [0,1]$. Значения весов могут задаваться пользователями, экспертами, алгоритмически и т. д. Структуру данных, представляющую информацию о пользователе u и объекте i назовем контентом пользователя $c_X(u)$ и контентом объекта $c_Y(i)$ соответственно.

Модель РС – это тройка

$$(c_X; c_Y; \text{П}), \quad (1)$$

где П – правило алгоритмического вычисления значения прогнозной функции $\overline{\rho}$.

Чтобы определить качество решения задачи, проводится тестирование, для которого исходное множество данных P разбивается на обучающее и тестовое множества P_0 и $P_\perp$ соответственно. Если $\rho(u, i) \in P_0$, будем обозначать такие объекты i_0. Если $\rho(u, i) \in P_\perp$, будем обозначать такие объекты $i_\perp$.

2 Коллаборативные модели

Рассмотрим коллаборативную фильтрацию [3-7], которая является одним из наиболее изученных [3], популярных [4] и успешных [5] правил вычисления П. РС, которые используют коллаборативную фильтрацию в качестве правила П, будем называть коллаборативными РС (далее КРС). Они делятся на два типа по фильтруемому множеству [6]: множеству пользователей или объектов. Будем называть первые субъектно-ориентированными (далее СОК), а последние – объектно-ориентированными (далее ОРС) [7].

Опишем теорию, на которой основаны коллаборативные П. Решение строится по обучающему множеству, а его качество определяется по тестовому. Обучающее множество выступает в роли информации прошедшего времени, тестовое – в роли информации будущего времени.

Правило П СОК основано на утверждении, которое гласит, что если в прошлом пользователи были близки по предпочтениям, то и в будущем они будут близки по предпочтениям. Во введенной терминологии данное утверждение примет следующий вид:

$$u_a \mathcal{R}_u u \text{ выполняется на } P_0 \Rightarrow u_a \mathcal{R}_u u \text{ выполняется на } P_\perp, \quad (2)$$

$\mathcal{R}_u$ – отношение близости пользователей. Выполнение отношения близости $\mathcal{R}_u$ между пользователями устанавливается СРС на основании значений характеристик пользователей. Характеристиками для СОК всегда выступают объекты, а значениями весов – значения $\rho(u, i) \in P_0$, которые были выставлены самими пользователями и характеризуют предпочтения пользователей. Для определения близости по предпочтениям используются так называемые меры близости

$$\delta_u: U \times U \to [0,1]: (1 - \delta_u(u, v)) \leq \varepsilon_0 \Leftrightarrow u \mathcal{R}_u v.$$

Пользователи, между которыми выполняется отношение близости, называются *соседями.*

Правило П СОК задается формулой

$$u \in U, (u_a \mathcal{R} u) \Rightarrow \left|\overline{\rho}(u_a, i_p) - \rho(u_a, i_p)\right| \leq \varepsilon_0, \quad (3)$$

$\overline{\rho}(u_a, i_p) = f(\{\rho(u, i_p)\})$. Правило П СОК говорит о том, что если пользователи u являются соседями для пользователя u_a, то оценки $\rho(u_a, i_p)$, $\rho(u, i_p)$ коррелируют, поэтому неизвестное значение $\rho(u_a, i_p)$ можно функционально определить по значениям $\{\rho(u, i_p)\}$, то есть прогнозная функция является функцией от значений оценок близости соседей.

Правило П ООК основано на утверждении: если пользователю нравится объект i, который близок по характеристикам к объекту j, то пользователю понравится объект j. Во введенной терминологии данное утверждение примет вид

$$(u_a \mathcal{R} i) \wedge (i \mathcal{R}_i j) \Rightarrow u_a \mathcal{R} j, \quad (4)$$

$\mathcal{R}_i$ – отношение близости объектов. Отношение близости $\mathcal{R}_i$ между объектами устанавливается РС на основании значений мер близости: $1 - \delta_i(i, j) \leq \varepsilon_0 \Leftrightarrow i \mathcal{R}_i j$, $\delta_i: I \times I \to [0,1]$ – мера близости объектов. Объекты, между которыми выполняется отношение близости, называются соседями.

При решения задачи $topN$ в ООК используется информация только о тех объектах, для которых известно, что $(u_a \mathcal{R} i_0), (u_a \mathcal{R} i_\perp)$, поэтому будем считать, что $P = \{\rho(u, i): u \mathcal{R} i\}$ для задачи $topN$.

Правило П ООК задается формулой

$$(i \mathcal{R}_i i_0) \Rightarrow (\overline{\rho}(u_a, i) = 0) \Rightarrow u_a \mathcal{R} i. \quad (5)$$

Значения $\overline{\rho}(u_a, i)$ задаются равными нулю, потому

что тогда объекты i будут близки активному пользователю при любом пороговом значении ε_0.

Правило вывода ООК говорит о том, что если существует объект i, являющийся соседом объекта i_0, то, следуя эвристическому утверждению, $u_a \mathcal{R} i$, так как $u_a \mathcal{R} i_0$ по принятому для задачи $topN$ виду исходного множества.

3 Проблемы применения коллаборативных моделей

3.1 Выполнение эвристических утверждений

Будем говорить, что РС *эффективна*, если ее правила вывода удовлетворяют некоторому критерию независимо от дополнительных ограничений или условий.

Реальные исходные данные обладают свойствами *динамики* и *неоднородности* [8]. Свойство динамики заключается в том, что множество исходных данных изменяется во времени, так как изменяются предпочтения пользователей, и мощность множеств U, I растет. Пусть выполняется $u_a \mathcal{R}_u u$ для P_0, но в силу динамики возможна ситуация, когда $1 - \delta_u(u_a, i) > \varepsilon_0$ для $P_\perp$. Тогда утверждение СОК (2) и, следовательно, правило П СОК (3) ложны в общем случае для любых исходных данных.

Свойство неоднородности заключается в том, что пользователи предпочитают различные объекты, не обязательно близкие по характеристикам, то есть их вкусы неоднородны: $(u_a \mathcal{R} i) \wedge (u_a \mathcal{R} j) \not\Rightarrow (i \mathcal{R}_i j)$. Тогда $(u_a \mathcal{R} i) \wedge (i \mathcal{R}_i j) \not\Rightarrow u_a \mathcal{R} j$, то есть утверждение ООК (4) и, следовательно, правило вывода ООК (5) ложны в общем случае для любых исходных данных. Таким образом, КРС не являются эффективными по критерию качества решения, так как оно зависит от выполнения эвристических утверждений, что, в свою очередь, зависит от свойств исходных данных.

Таким образом, КРС не являются эффективными по критерию качества решения, так как оно зависит от выполнения эвристических утверждений, что, в свою очередь, зависит от свойств исходных данных.

3.2 Достаточные условия качественного решения

Отношение близости обладает следующими свойствами: рефлексивность, симметричность, транзитивность. Выполнение свойства *транзитивности* отношения близости зависит от выбора функции, используемой в качестве меры близости, и значения порогового значения ε_0.

Пусть правила вывода П СОК (3) и ООК (5) истинны (то есть выполняются эвристические утверждения). Рассмотрим условия, которые влияют на качество решения.

Достаточным условием, при котором СОК гарантирует получение качественного решения задачи прогнозирования, является транзитивность отношения близости на кластере соседей $\mathcal{N}_U = \{u : u_a \mathcal{R} u\}$, который строится для решения задачи: $\forall u_1, u_2 \in \mathcal{N}_U : (u_1 \mathcal{R}_u u_a) \wedge (u_2 \mathcal{R}_u u_a) \Rightarrow u_1 \mathcal{R}_u u_2$. Назовем это условие *условием 1*.

Достаточным условием, при котором ООК гарантирует получение качественного решения задачи $topN$, является транзитивность отношения близости на объединении обучающего, тестового и результирующего множеств:

$$(i \mathcal{R}_i j) \wedge (i \mathcal{R}_i k) \Rightarrow (j \mathcal{R}_i k), i, j, k \in I_0 \cup I_{topN} \cup I_\perp,$$
$$I_\perp = \{i_\perp, I_0 = \{i_0\}.$$

Назовем его *условием 2*.

Выполнение достаточных условий зависит от того, какое значение выбрано в качестве порогового значения ε_0, и функции, использующейся в качестве меры близости. К примеру, если $\delta_i = \cos$ и $\varepsilon_0 = 0{,}49$, то транзитивность не гарантируется; коэффициент корреляции Пирсона [6], являющийся традиционной мерой близости СОК, не обладает свойством транзитивности [9].

Если эвристические утверждения выполняются, то правила вывода П гарантируют получение качественного решения, если выполняются достаточные условия, что зависит от разработчиков системы.

Проблемы, описанные в разделах 2.1 и 2.2, подтверждены на практике и продемонстрированы ниже в Разделе 4, в Таблицах 1 и 2.

3.3 Масштабируемость

Стандартные алгоритмы решений КРС обладают следующими асимптотическими сложностями [10]: $O(|I|^2)$ для задачи $topN$, $O(|U|)$ для задачи прогнозирования. Учитывая огромную мощность множеств U, I, такие асимптотические сложности приводят к проблеме масштабируемости КРС [10].

4 Нечеткая контентная модель

4.1 Описание

В нечеткой контентной модели будем представлять контент в виде нечеткое подмножества множества характеристик [11]: $\{(c | w_M(m, c))\}$, где c – характеристика пользователя или объекта, $m \in M$ – множество пользователей или объектов, w_M – характеристическая функция принадлежности. Для СОК и ООК контент пользователя представляется в виде нечеткого множества вида $\{(i | 1 - \rho(u, i))\}$. Между пользователями и объектам введем расстояния ρ_u и ρ_i соответственно как обобщенное расстояние Хэмминга, которое обладает метрическими свойствами.

При представлении контентов в виде нечетких множеств определим нечеткое отображение $\Psi : U \to I$, характеристическая функция которого задана следующей формулой:

$$\nu_\Psi(y) = \max_{x \in X} \min\{\delta_c(x, y); w_U(u, x)\}, \quad (6)$$

и расстояние между пользователем и объектом

$$\overline{\rho}(u, i) = \rho_i(\Psi(u), i). \quad (7)$$

Функция $\delta_c: X \times Y \to [0,1]$ – это функция сходства характеристик пользователей и объектов, задание которой необходимо для построения отображения Ψ. Эта функция может быть определена разработчиками РС, экспертами, алгоритмически и т. д. Будем говорить, что оценка сходства δ_c задана аккуратно, если выполняется неравенство

$$|\rho(u,i) - \overline{\rho}(u,i)| \le \varepsilon_0 \quad (9)$$

Нечеткое правило вычисления Π_f заключается в задании оценки сходства δ_c, нечеткого отображения Ψ (6) и вычислении расстояния между пользователем и объектом, определенного формулой (7):

$$\Pi_f = \{\delta_c, (6), (7)\}. \quad (10)$$

Нечеткая контентная модель – это модель, которая задается следующей тройкой:

$$(c_X; c_Y; \Pi \in \{\Pi_f, \Pi_{\text{СОК}}, \Pi_{\text{ООК}}\}). \quad (11)$$

4.2 Нечеткая модель как эффективное расширение коллаборативной модели

Утверждение 1: нечеткая контентная модель (11) является эффективным расширением СОК по критерию качества.

Утверждение 1 следует из того, что СОК – частный случай модели (11) при использовании $\Pi = \Pi_{\text{СОК}}$. Расширение эффективно по критерию качества, так как выполняется условие 1. Покажем, что это верно: введем следующее дополнительное условие при составлении кластера соседей – $\mathcal{N}_U = \{u: \rho_u(u_a, u) \le \varepsilon_0/2\}$. Покажем, что выполняется достаточное условие. Напомним, что оно заключается в выполнении свойства транзитивности отношения близости $\mathcal{R}_u$ на кластере соседей: $\forall u, v \in \mathcal{N}_U$ верно, что $(u_a \mathcal{R}_u u) \wedge (u_a \mathcal{R}_u v)$.

Так как функция ρ_u обладает метрическими свойствами, то $\rho_u(u, v) \le \rho_u(u_a, u) + \rho_u(u_a, v)$. По дополнительному условию $\rho_u(u_a, u) \le \varepsilon_0/2$, $\rho_u(u_a, v) \le \varepsilon_0/2$, поэтому $\rho_u(u, v) \le \varepsilon_0 \Rightarrow u \mathcal{R}_u v$.

Утверждение 2: нечеткая контентная модель (11) является эффективным расширением ООК по критерию качества.

Утверждение 2 следует из того, что ООК – частный случай модели (11) при использовании $\Pi = \Pi_{\text{ООК}}$. Расширение эффективно по критерию качества, так как выполняется условие 2 при $\varepsilon_0 = 0$. Покажем, что выполняется условие 2. Напомним, что оно заключается в выполнении свойства транзитивности отношения близости $\mathcal{R}_i$ на множестве $I_0 \cup I_\perp \cup I_{topN}$. Покажем, что $(i_0 \mathcal{R}_i i) \wedge (i_0 \mathcal{R}_i i_\perp) \Rightarrow i_\perp \mathcal{R}_i i$.

Отношение $i_0 \mathcal{R}_i i_\perp$ выполняется по эвристическому утверждению ООК (4), отношение $i_0 \mathcal{R}_i i$ выполняется по построению решения. Так как функция ρ_i обладает метрическими свойствами, то $\rho_i(i, i_\perp) \le \rho_i(i_0, i) + c_i(i_\perp, i_0)$. По дополнительному условию $\rho_i(i_0, i) = 0$, так как выполняется $i_0 \mathcal{R}_i i_\perp$, то $\rho_i(i_0, i_\perp) \le \varepsilon_0$. Поэтому $\rho_i(i, i_\perp) \le \varepsilon_0$, то есть выполняется отношение $i \mathcal{R}_i i_\perp$.

Таким образом, правила вывода $\Pi_{\text{СОК}}, \Pi_{\text{ООК}}$ в представлении контентов в виде нечетких подмножеств и при использовании метрических расстояний обладают большей эффективностью по критерию качества решения, так как выполняются достаточные условия 1 и 2, и поэтому контентная нечеткая модель является эффективным расширением по критерию качества. Данный вывод подтверждается практическими результатами (см. таблицы 1 и 2).

4.3 Применение нечеткого правила вывода для решения задач

Определим решения в нечеткой контентной модели при использовании Π_f.

Задача $topN$ может быть решена при помощи *линейного поиска* объектов, таких, что $\overline{\rho}(u_a, i) \le \varepsilon_0$. Асимптотическая сложность такого алгоритма равна $O(|I|)$.

Для решения задачи прогнозирования нужно всего лишь рассчитать значение $\overline{\rho}(u_a, i_p)$, поэтому асимптотическая сложность такого решения равна $O(C)$.

Если оценка сходства δ_c *задана аккуратно*, то решения задач контентной нечеткой модели *эффективны* по критерию качества, что будет продемонстрировано на разделе 4. Точность задания δ_c зависит только от разработчиков, но не от свойств исходных данных или дополнительных условий, как в случае с КРС.

Асимптотические сложности алгоритмов решений при использовании правила вывода Π_f на порядок меньше по сравнению со сложностями КРС, поэтому представленная модель более эффективна по критерию масштабируемости, чем КРС. Каждый раз, когда производится вычисление $\overline{c}$, производятся отображение Ψ и расчет δ_c. Сложности вычислений отображения Ψ и δ_c зависят от мощности контента (которое, как правило, значительно меньше мощности множеств пользователей и объектов) и от того, как была задана δ_c разработчиками, поэтому эти сложности не учтены в расчетах, представленных ниже.

Приведенные значения асимптотических сложностей показывают, что контентная нечеткая модель является эффективным расширением КРС по критерию масштабируемости.

5 Практические результаты

Для получения практических результатов было разработано программное обеспечение, которое реализует ООК, СОК и нечеткую контентную РС. С помощью ООК решалась задача $topN$, с помощью СОК – задача прогнозирования. С помощью нечеткой РС решались обе задачи.

Тестирование проводилось на множестве данных, сформированных компанией MovieLens. Множество

данных имеет следующие характеристики:

- $|I|$=10000 – объектами множества являются фильмы, численность которых равна 10000;
- $|Y| = 18$ – множество характеристик объектов состоит из 18 кинематографических жанров;
- $|U| = 670$ – число пользователей данного множества равно 671; пользователи являются реальными людьми, которые проставляли оценки близости различным объектам.

Для решения задач $topN$ и прогнозирования в ООК и СОК соответственно были использованы стандартные алгоритмы и подходы [6, 12]. При решении задачи $topN$ в ООК использовалась мера сходства косинус, при решении задачи прогнозирования в СОК – коэффициент корреляции Пирсона. Те же алгоритмы были применены при решении задач в нечеткой контентной модели, но при этом использовались расстояния ρ_i и ρ_u. Пороговое значение ε_0 было принято равным 0,1.

Чтобы применить Π_f, была задана функция δ_c на основании эвристического предположения о том, что между оценкой пользователя и жанрами объектов существует корреляция:

$$\delta_c(i,y) = (|like_y| - |dislike_y|)/|P_u|.$$

Если $\delta_c(i,y) < 0$, то $\delta_c(i,y) = 0{,}0001$;

$$like_y = \{i: (\rho(u,i) \le \varepsilon_0) \wedge w_U(i,y) \neq 0\},$$

$$dislike_y = \{i: (\rho(u,i) > \varepsilon_0) \wedge w_U(i,y) \neq 0\},$$

$$P_u = \{i: (\rho(u,i) \neq \perp)\}.$$

Такое эвристическое предположение верно не для всех пользователей, так как их вкусы могут быть неоднородными. Поэтому для некоторых пользователей функция δ_c задана аккуратно, а для некоторых – нет.

Стандартно при проведении тестирования данные о пользователе случайно разбивались в следующем отношении: 80% – обучающее множество, 20% – тестовое. Обозначим такое разбиение цифрой 1. Помимо стандартного разбиения использовались и другие специально сформированные разбиения 2 и 3. Разбиение 2 составлено так, что обучающее множество состоит из таких объектов i, для которых выолняется отношение $\mathcal{R}_i$, тестовое множество состоит из таких объектов j, для которых отношение $i\,\mathcal{R}_i\,j$ не выполняется. Такое разбиение создано для того, чтобы подтвердить или опровергнуть влияние свойства неоднородности данных на эффективность по критерию качества. Разбиение 3 составлено так же, как и стандартное разбиение, но в нем участвуют только те пользователи, для которых функция δ_c задана аккуратно.

Эффективность решений задач по критерию качества определяется усредненными по числу тестов (равному 1000 для каждой задачи, разбиению и модели) значениями функций. Эффективность решения задачи $topN$ по критерию качества оценивалась значениями функций точность (P), точность по списку длины L, средняя точность, NDCG. В результате тестирования среднее значение этих функций мало отличалось, поэтому в Таблице 1 приведены только значения точности. Большее значение точности свидетельствует о то, что решение более эффективно. Эффективность решения задачи прогнозирования по критерию качества оценивалась значениями функций MAE, NMAE, RMSE, меньшее значение которых говорит о более эффективном решении.

Таблица 1

№	Модель/Правило вычисления	Разбиение	P
1	ООК/ $\Pi_{ООК}$	1	0.32
2	ООК/ $\Pi_{ООК}$	2	0,24
3	Нечеткая контентная / $\Pi_{ООК}$	1	0,55
4	Нечеткая контентная / $\Pi_{ООК}$	2	0,53
5	Нечеткая контентная/ Π_f	1	0.39
6	Нечеткая контентная/ Π_f	2	0,36
7	Нечеткая контентная/ Π_f	3	0,81

Прокомментируем данные таблиц 1 и 2. Результаты 1 эффективнее результатов 2 и результаты 3 эффективнее результатов 4, что подтверждает теоретические выводы о влиянии свойства неоднородности на эффективность по критерию качества при применении ООК. Разбиение 2 задано так, что свойства неоднородности влияют на эффективность решения, так как между объектами обучающего и тестового множеств не выполняется отношение сходства, в результате чего нарушается утверждение ООК (4). Разбиение 2 увеличивает вероятность того, что утверждение СОК (5) может быть неверным, поэтому результаты 1 и 3 эффективней результатов 2 и 4 Таблицы 2.

Таблица 2

№	Модель/ Правило вычисления	Разбиение	MAE	NMAE	RMSE
1	ООК/ $\Pi_{СОК}$	1	0.14	0.23	0.19
2	ООК/ $\Pi_{СОК}$	2	0.16	0.26	0.21
3	Нечеткая контентная / $\Pi_{СОК}$	1	0.08	0.19	0.13
4	Нечеткая контентная / $\Pi_{СОК}$	2	0.10	0.17	0.18
5	Нечеткая контентная/ Π_f	1	0.14	0.23	0.21
6	Нечеткая контентная/ Π_f	2	0.16	0.26	0.22
7	Нечеткая контентная/ Π_f	3	0.05	0.04	0.1

Результаты 3 и 4 эффективнее результатов 1 и 2, что подтверждает вывод о том, что нечеткая контентная модель является эффективным расширением, так как в ней выполняются достаточные условия 1 и 2. Эти же результаты подтверждают выводы о влиянии меры сходства на эффективность ООК и СОК по критерию качества.

Результаты 7 эффективнее результатов 3–6, так как для разбиения 7 функция δ_c задана аккуратно. Результаты 7 эффективнее результатов 5 и 6, так как для 5 и 6 в общем случае δ_c не задана аккуратно, и поэтому же 5 и 6 не эффективнее 3 и 4. Результаты 5 эффективнее 6, так как функция δ_c задавалась на основании данных обучающего множества, поэтому свойство неоднородности влияет на аккуратность функции так же, как и на эффективность КРС по критерию качества. Использование Π_f может быть неэффективным, если о пользователях известна только та информация, которая принадлежит исходному множество P. В такой ситуации эффективнее использовать нечеткую модель Π_{OOK} или Π_{COK}. Для задания функции δ_c можно использовать информацию, которая никак не зависит от мощности и свойств исходных данных, и тогда решения задач в нечеткой контентной модели не будут зависеть от свойств исходных данных. Такой информацией может выступать, к примеру, контекстная информация [13].

6 Заключение

Нечеткая контентная модель РС, представленная в настоящей работе, является эффективным расширением КРС по критериям качества решений и масштабируемости.

Литература

[1] Resnick, P., Varian, H.R.: Recommender systems. Communications of the ACM, 40 (2), pp. 56-58 (1997)

[2] Goldberg, D., Nichols, D., Oki, B.M., Terry, D.: Using collaborative filtering to weave an information tapestry. Communications of the ACM, 35 (12), pp. 61-70 (1992)

[3] Asanov, D.: Algorithms and Methods in Recommender Systems. Berlin Institute of Technology. https://www.snet.tu-berlin.de/fileadmin/ fg220/courses/SS11/snet-project/ recommender-systems_asanov.pdf

[4] Yao, W., Xudong, L., Min, X., Ester, M., Qing, Y.: CCCF: Improving Collaborative Filtering via Scalable User-Item Co-Clustering. WSDM '16 Proc. of the Ninth ACM Int. Conf. on Web Search and Data Mining, pp. 73-82 (2016)

[5] Hu, R., Pu, P.: Using personality information in collaborative filtering for new users. Recommender Systems and the Social Web, pp. 17-24 (2010)

[6] Su, X., Khoshgoftaar, T.: A survey of collaborative filtering based social recommender systems. Computer Communications, 41, pp. 1-10 (2014)

[7] Wang Jun: Unifying user-based and item-based collaborative filtering approaches by similarity fusion. SIGIR'06 Proc. of the 29th Annual International ACM, pp. 501-508 (2006)

[8] Посыпанова, О.: Экономическая психология: психологические аспекты поведения потребителей. Калуга: Изд-во Калужского государственного университета им. К.Э. Циолковского, 296 с. (2012)

[9] Castro Sotos, A., Vanhoof, S., Van den Noortgate, W., Onghena, P.: The non-transitivity of Pearson's correlation coefficient: an educational perspective. Proc. of the 56th Session of the ISI, 62, pp. 4609-4613 (2007)

[10] Linden, G., Smith, B., York, J.: Amazon.com Recommendations Item-to-Item Collaborative Filtering. Internet Computing, IEEE, 7, pp. 76-80 (2003)

[11] Амелькин, С.А., Понизовкин, Д.П.: Математическая модель задачи topN для контентных рекомендательных систем. Изв. МГТУ МАМИ, 2, сс. 26-31 (2013)

[12] Deshpande, M., Karypis, G.: Item-Based Top-N Recommendation Algorithms. ACM Transactions on Information Systems, 22 (1), pp. 143-177 (2004)

[13] Adomavicius, G., Tuzhilin, A.: Context-aware recommender systems. Conference: Proc. of the 2008 ACM Conference on Recommender Systems, RecSys 2008, Lausanne, Switzerland, October 23–25, 2008. doi: 10.1007/978-1-4899-7637-6_6

Распределенные вычисления

Distributed computing

Моделирование задержек передачи информации в вычислительном кластере для мониторинга коммуникационной среды

© А.И. Майсурадзе © В.Д. Козлов

Московский государственный университет имени М.В. Ломоносова,
Москва, Россия

maysuradze@cs.msu.ru kozlov2.volodia2@gmail.com

Аннотация. Эффективное использование современных вычислительных кластеров опирается не только на характеристики составляющих их узлов, но и на характеристики коммуникационной среды. Чтобы проверять работоспособность коммуникационной среды и динамически планировать расписание заданий, существуют различные подходы. В данной работе рассмотрен подход, опирающийся на предварительный сбор информации о задержках передачи сообщений и их последующий анализ. Такой сбор занимает много времени и порождает большое количество первичной информации. Требуются модели задержек, которые позволяют существенно ускорить сбор данных и сократить объем хранимой информации. В работе предложены такая модель и методы её настройки, которые сочетают высокое качество и скорость.

Ключевые слова: вычислительный кластер, задержка передачи сообщений, анализ коммуникационной среды, сбор данных, настройка модели.

Modeling Message Passing Delays in a Computer Cluster to Monitor its Network

© A. Maysuradze © V. Kozlov

Lomonosov Moscow State University
Moscow, Russia

kozlov2.volodia2@gmail.com maysuradze@cs.msu.ru

Abstract. The effective use of a modern computer cluster relies not only on the characteristics of its nodes, but its communication environment as well. There are different approaches to monitor communication environment and dynamically schedule tasks. In the paper, we consider an approach based on the preliminary collection of data on delays of the message passing and their subsequent analysis. This collection takes a lot of time and generates a large amount of raw data. Delay models are required that can significantly speed up data collection and reduce the amount of stored information. We proposes and study such a model and methods of its learning, which combine high quality and speed.

Keywords: computer cluster, message passing delay, network analysis, data collection, model learning.

1 Введение

Современные распределённые вычислительные системы состоят из тысяч и десятков тысяч процессоров. Увеличение числа процессоров ведёт к усложнению коммуникационной среды и росту накладных расходов на обмен информацией между вычислительными устройствами. Эффективность современных многопроцессорных систем зависит не только от характеристик отдельных вычислительных устройств, но и от характеристик коммуникационной среды.

Одним из ключевых инструментов разработки параллельных приложений для многопроцессорных систем является библиотечная реализация стандарта MPI (Message Passing Interface). При использовании технологии MPI программа разделяется на процессы, взаимодействующие посредством обмена сообщениями. Информация о задержках, возникающих при передаче сообщений, может быть использована для повышения эффективности работы вычислительной системы, в частности, решения задач динамического планирования выполнения параллельных программ, а также для диагностики коммуникационной среды. Таким образом, возникает потребность в моделировании задержек. При этом модель должна не только полной, но достаточно компактной, чтобы обеспечить хранение

Труды XIX Международной конференции «Аналитика и управление данными в областях с интенсивным использованием данных» (DAMDID/ RCDL'2017), Москва, Россия, 10–13 октября 2017 года

и использование в реальном времени информации о задержках при передаче сообщений для каждой пары вычислительных узлов.

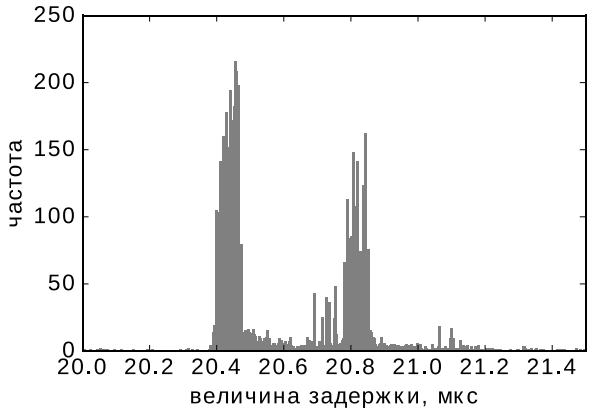


Рисунок 1 Пример картины задержек при передаче сообщений. Суперкомпьютер BlueGene/P

Величины задержек зависят от множества факторов, специфичных для разных вычислительных систем и меняющихся со временем, учёт которых при моделировании задержек требует анализа программного и аппаратного обеспечения на всех уровнях сетевого протокола, что возможно лишь для самых простых архитектур. В связи с этим начали активно развиваться системы MPI-тестирования коммуникационной среды [13]. Поскольку на практике размеры вычислительных кластеров не позволяют хранить выборки задержек для всех пар вычислительных узлов, для описания используются некоторые эмпирические статистики, вычисленные по выборкам величин задержек, которые могут не отражать в полной мере структуры задержек. В качестве альтернативы предлагается стохастическая модель, в которой неконтролируемые факторы рассматриваются как скрытые параметры, а задержки – как случайные величины с некоторыми распределениями. Такая модель одновременно описывает картину задержек более полно, чем набор статистик, и позволяет хранить информацию в сильно сжатом виде – всего несколько чисел – параметров модели вместо выборки.

Проведенные ранее исследования задержек в локальных сетях и интернете [5, 10, 11] показали, что величины задержек хорошо описываются трёхпараметрическим гамма- или логнормальным распределением. В коммуникационных средах вычислительных кластеров, однако, наблюдаются следующие особенности [6]:

- распределение задержек является многомодальным;
- в данных много повторов и мало уникальных значений.

На Рис. 1 приведена картина задержек в коммуникационной среде суперкомпьютера BlueGene/P, на которой явно видны указанные особенности. Исходя из этого, в работе [6] в качестве модели задержек предложено использовать смесь трёхпараметрических логнормальных распределений. Однако проблемы возникают даже при параметрическом восстановлении одного компонента такой смеси. Подробнее об этих проблемах сказано ниже.

Работа посвящена разработке и исследованию специализированных методов восстановления трёхпараметрических логнормальных распределений по конечным выборкам задержек передачи информации в коммуникационной среде суперкомпьютера. Статья устроена следующим образом. В разделе 2 введены используемые основные определения Разделы 3, 4 и 5 посвящены обзору существующих методов оценки параметров трёхпараметрического логнормального распределения (метод максимального правдоподобия, метод моментов, метод L-моментов и методы минимизации расстояния). В разделе 6 описаны данные, использованные в вычислительном эксперименте (модельные и реальные). Раздел 7 посвящён сравнению методов оценивания параметров, рассмотренных в разделах 3, 4 и 5, на синтетических и реальных данных.

2 Основные обозначения и определения

Трёхпараметрическое логнормальное распределение (3LN распределение) – это абсолютно непрерывное одномерное распределение, функция плотности вероятности которого выражается формулой

$$p(x;\,\gamma,\mu,\sigma) = \begin{cases} \dfrac{1}{\sqrt{2\pi}\sigma(x-\gamma)}\exp\left(-\dfrac{(\ln(x-\gamma)-\mu)^2}{2\sigma^2}\right), & x \geq \gamma, \\ 0, & x < \gamma. \end{cases}$$

Функция распределения 3LN может быть записана в виде $F(x;\gamma,\mu,\sigma) = \Phi\left(\frac{\ln(x-\gamma)-\mu}{\sigma}\right)$, где $\Phi(x)$ – функция распределения стандартного нормального закона [4]. Основные моментные характеристики распределения указаны в таблице 1.

Набор параметров γ,μ,σ будем обозначать θ. Будем обозначать случайную выборку длины n $X^n = (X_1,\ldots,X_n)$, её реализацию – $x^n = (x_1,\ldots,x_n)$, k-ю порядковую статистику и её реализацию – $X_{(k)}$ и $x_{(k)}$ соответственно.

3 Метод максимума правдоподобия и его модификации

Наиболее популярным подходом к параметрической оценке плотности распределения является метод максимального правдоподобия (ММП). В качестве меры адекватности распределения $F(\cdot,\theta)$ данным x^n используется функция правдоподобия $L(\theta) = p(x^n;\,\theta)$ – совместная плотность вероятности объектов выборки. Полагается, что чем больше значение функции правдоподобия, тем лучше модель описывает данные. Оценки максимального правдоподобия для многих задач оказываются состоятельными, асимптотически нормальными и асимптотически эффективными.

Для семейства 3LN распределений логарифм функции правдоподобия имеет вид

$$\ln L(\theta) = -\frac{n}{2}\ln 2\pi - n\ln\sigma - \sum_{i=1}^{n}\ln(x_i-\gamma) - \sum_{i=1}^{n}\left(\frac{(\ln(x_i-\gamma)-\mu)^2}{2\sigma^2}\right),$$

причём выражение имеет смысл только при $\gamma < x_{(1)}$. Для неё можно выписать необходимые условия экстремума:

$$\begin{cases} \frac{\partial \ln L}{\partial \gamma} = \sum_{i=1}^{n}\frac{1}{x_i-\gamma}\left(1+\frac{\ln(x_i-\gamma)-\mu}{\sigma^2}\right) = 0, \\ \frac{\partial \ln L}{\partial \mu} = \sum_{i=1}^{n}\frac{\ln(x_i-\gamma)-\mu}{\sigma^2} = 0, \\ \frac{\partial \ln L}{\partial \sigma} = \sum_{i=1}^{n}\frac{1}{\sigma}\left(-1+\frac{(\ln(x_i-\gamma)-\mu)^2}{\sigma^2}\right) = 0. \end{cases}$$

Метод максимального правдоподобия с успехом используется во многих задачах статистики, однако его применимость для оценки параметров 3LN распределения оказывается под вопросом. Показано [8], что для любой выборки x^n функция правдоподобия $L(\theta)$ не ограничена, и существуют траектории в пространстве параметров (γ,μ,σ), сходящиеся к $(x_{(1)},-\infty,+\infty)$, при движении вдоль которых $L(\theta)$ сходится к $+\infty$, при этом в точке $(x_{(1)},-\infty,+\infty)$ $L(\theta)$ принимает значение 0. Таким образом, возникает потребность в использовании иных методов оценки параметров 3LN распределения.

Несмотря на общую неограниченность функции правдоподобия $L(\theta)$, если элементы выборки принимают достаточно много различных значений, «вблизи» истинных значений параметров функция правдоподобия имеет локальный максимум [8]. Это приводит к идее использования так называемых локальных оценок максимального правдоподобия. В работе [7] показано, что такие оценки для 3LN распределения обладают хорошими асимптотическими свойствами.

Для поиска оценок локального максимума предлагается использовать необходимые условия экстремума логарифмической функции правдоподобия [3]. Параметры μ и σ^2 выражаются как функции параметра γ:

$$\begin{cases} \mu(\gamma) = \frac{1}{n}\sum_{i=1}^{n}\ln(x_i-\gamma), \\ \sigma^2(\gamma) = \frac{1}{n}\sum_{i=1}^{n}(\ln(x_i-\gamma)-\mu)^2, \end{cases}$$

после чего оценка параметра γ получается из уравнения

$$\lambda(\gamma) = \sum_{i=1}^{n}\frac{1}{x_i-\gamma}\left(1+\frac{\ln(x_i-\gamma)-\mu(\gamma)}{\sigma^2(\gamma)}\right) = 0.$$

4 Общий метод моментов

При оценке параметров с использованием метода моментов на распределение $F(\cdot;\theta)$ накладывается последовательность ограничений типа равенства, образующая систему уравнений вида $g_i(\theta) = h_i(X^n), i=\overline{1,k}$, где функции $g_i(\theta)$ характеризуют теоретическое распределение, а $h_i(X^n)$ являются их выборочными оценками, как правило, несмещёнными или хотя бы асимптотически несмещёнными.

Таблица 1 Основные моменты 3LN распределения с параметрами $\boldsymbol{\gamma}$, $\boldsymbol{\mu}$ и $\boldsymbol{\sigma}$ ($\boldsymbol{\beta} = \exp\boldsymbol{\mu}, \boldsymbol{\omega} = \exp\boldsymbol{\sigma^2}$)

Математическое ожидание $\mathbb{E}$	$\gamma + \beta\sqrt{\omega}$
Дисперсия $\mathbb{D}$	$\beta^2\omega(\omega-1)$
Коэффициент асимметрии α_3	$\sqrt{\omega-1}(\omega+2)$
Коэффициент эксцесса α_4	$\omega^4+2\omega^3+3\omega^2-6$

Общий метод моментов применялся для оценки параметров 3LN распределения в связи с указанными выше проблемами, возникающими при использовании метода максимума правдоподобия [4]. В качестве функций $g_i(\theta), i=1,2,3$, использовались математическое ожидание, дисперсия и коэффициент асимметрии (таблица 1), в качестве $h_i(X^n), i=1,2,3$, – их выборочные оценки [4]. Итогом является система уравнений

$$\begin{cases} \gamma+\beta\sqrt{\omega} = \frac{1}{n}\sum_{i=1}^{n}x_i, \\ \beta^2\omega(\omega-1) = \frac{1}{n-1}\sum_{i=1}^{n}(x_i-\overline{x})^2, \\ \sqrt{\omega-1}(\omega+2) = \frac{\frac{n}{(n-1)(n-2)}\sum_{i=1}^{n}(x_i-\overline{x})^3}{(\frac{1}{n-1}\sum_{i=1}^{n}(x_i-\overline{x})^2)^{\frac{3}{2}}}. \end{cases}$$

Третье уравнение не содержит переменных γ и β и имеет вид $\omega^3+3\omega^2-(4+a^2)=0$. Если $a^2>0$, уравнение имеет единственное решение, большее 1, которое вычисляется по формуле

$$\omega = 1+\left(\sqrt[3]{\frac{(a_3+4)^2+a_3}{2}}-\sqrt[3]{\frac{(a_3+4)^2-a_3}{2}}\right)^2.$$

Оценки для $\sigma=\sqrt{\ln\omega}$, $\mu=\ln\beta$ и γ получаются аналитически.

В работе [2] для оценки параметров 3LN распределения предложено использовать метод L-моментов. Теоретическим L-моментом порядка r для распределения $F(x)$ называется величина

$$\lambda_r = \frac{1}{r}\sum_{k=0}^{r-1}(-1)^k\binom{r-1}{k}\mathbb{E}X_{(r-k)},$$

то есть L-момент представляет собой линейную комбинацию математических ожиданий порядковых статистик распределения специального вида. Выборочный L-момент порядка $r \le n$ определяется как

$$l_r = \binom{n}{r}^{-1}\sum_{1\le i_1\le i_2\le\cdots\le i_r\le n}\frac{1}{r}\sum_{k=0}^{r-1}(-1)^k\binom{r-1}{k}x_{(i_{r-k})}.$$

Эти статистики являются несмещёнными оценками теоретических L-моментов. Метод L-моментов имеет ряд преимуществ по сравнению с «обычным» методом моментов: L-моменты однозначно определяют параметры, устойчивы к выбросам в данных, а при малых размерах выборки зачастую дают более качественные оценки, чем метод максимального правдоподобия [9].

Для 3LN распределения можно выписать следующую систему уравнений [9]:

$$\begin{cases} \gamma + \exp\left(\mu + \dfrac{\sigma^2}{2}\right) = l_1, \\ \exp\left(\mu + \dfrac{\sigma^2}{2}\right)\operatorname{erf}\left(\dfrac{\sigma}{2}\right) = l_2, \\ \dfrac{6}{\sqrt{\pi}}\dfrac{\int_0^{\frac{\sigma}{2}}\operatorname{erf}\left(\frac{x}{\sqrt{3}}\right)\exp(-x^2)\,dx}{\operatorname{erf}\left(\frac{\sigma}{2}\right)} = \dfrac{l_3}{l_2}, \end{cases}$$

где erf – функция ошибок. Для этой системы можно найти приближённое решение [2]:

$$\begin{aligned} z &= \sqrt{(8/3)}\Phi^{-1}\left(\frac{1+l_3/l_2}{2}\right), \\ \sigma &\approx 0{,}999281z - 0{,}006118z^3 + 0{,}000127z^5, \\ \mu &= \ln\left(\frac{l_2}{\operatorname{erf}\left(\frac{\sigma}{2}\right)}\right) - \frac{\sigma^2}{2}, \\ \gamma &= l_1 - \exp\left(\mu + \frac{\sigma^2}{2}\right). \end{aligned}$$

5 Метод минимизации расстояния

В методе минимизации расстояний мерой соответствия модели данным служит некоторым образом выбранное расстояние $d[\cdot,\cdot]$ между теоретическим и эмпирическим распределениями данных. Полагается, что чем меньше расстояние, тем лучше модель описывает данные. Для непрерывных распределений расстояние обычно берётся между функцией распределения модели $F(x;\theta)$ и эмпирической функцией распределения $F_n(x) = \frac{1}{n}\sum_{i=1}^{n}\mathbb{I}[x > x_{(i)}]$ [1]. Следует отметить, что термин «расстояние» используется условно: функционал d может быть даже несимметричен, обычно от него требуются только неотрицательность и равенство нулю только в случае равенства распределений. Оценкой минимального расстояния θ_0 называется $\theta_0 = \underset{\theta\in\Theta}{\operatorname{argmin}}\, d[F(x;\theta), F_n(x)]$. Одним из наиболее привлекательных свойств оценок минимального расстояния является их робастность, то есть устойчивость к возмущениям в данных [1].

Применительно к задаче оценки параметров 3LN распределения в работе [6] отмечалось, что методы минимизации расстояния, как правило, оказываются предпочтительнее других методов: они дают более точные оценки параметров, чем другие методы, в частности, метод максимального правдоподобия, и они не страдают от проблем со сходимостью оптимизационной процедуры. Несмотря на эти положительные свойства, до нас никто подробно не исследовал применение методов минимизации расстояния к задаче оценки параметров 3LN распределения.

В работе рассматриваются расстояния Колмогорова – Смирнова, Крамера – фон Мизеса и Андерсона – Дарлинга [14].

6 Модельные и реальные данные

Ниже нам предстоит настраивать и сравнивать отобранные методы оценивания параметров. Для этого мы использовали модельные и реальные данные из рассматриваемой предметной области.

Таблица 2 Параметры модельных распределений

	θ_1	θ_2	θ_3	θ_4	θ_5	θ_6	θ_7
γ	3	10	16	10	10	10	10
μ	3	3	3	2	4	3	3
σ	0.23	0.23	0.23	0.23	0.23	0.1	0.35

В качестве модельных данных использовались выборки из 3LN распределения с известными параметрами θ. По результатам анализа задержек при передаче сообщений в локальной сети и интернете [10] мы выбрали 7 наборов параметров γ, μ, σ, для каждого из них генерировали несколько выборок. Значения параметров модельных данных приведены в таблице 2. Поскольку для модельных данных параметры известны, можно сравнивать полученные оценки с истинным значением параметра и исследовать их статистические свойства. Следует отметить, однако, что такие модельные данные относятся к иной, хотя и смежной предметной области и не обладают особенностями рассматриваемой задачи (см. введение).

Также мы проанализировали работу методов на реальных данных о задержках в коммуникационной сети вычислительной системы BlueGene/P. Эти данные отвечают предметной области и имеют особенности, указанные в разделе 1. Для сбора данных использовалась утилита `network_test2` из пакета PARUS [12]. В силу многомодальности данные для анализа выделялись из выборки вручную.

7 Сравнение методов оценки параметров

С целью сравнения описанных выше методов оценки параметров распределения мы провели

тестирование на модельных данных. Для каждого набора параметров, указанных в таблице 2, было сгенерировано по 100 выборок длиной 10000 каждая. Для каждой выборки производилось оценивание параметров всеми описанными выше методами. Таким образом, для каждого модельного набора параметров и каждого метода оценивания мы получили по 100 оценок этих параметров. Для сравнения методов мы использовали следующие характеристики:

- среднеквадратичная ошибка для каждого из параметров γ, μ и σ: $\frac{1}{n}\sum_{i=1}^{n}(\theta_i - \theta)^2$, где n – число оценок ($n = 100$), θ_i – оценка параметра γ, μ или σ по выборке, θ – истинное значение этого параметра;
- среднее время работы метода на выборках (в секундах).

Результаты тестирования методов на модельных данных приводятся в таблицах 3–6. Используемые сокращения: К-С, К-фон М, А-Д – расстояния Колмогорова – Смирнова, Крамера – фон Мизеса и Андерсона – Дарлинга.

Таблица 3 Среднеквадратичная ошибка оценок параметра $\boldsymbol{\gamma}$

	ММП	моменты	L-моменты
θ_1	0.47	0.83	0.66
θ_2	0.48	0.86	0.63
θ_3	0.37	0.69	0.51
θ_4	0.61	0.13	0.09
θ_5	3.89	9.10	6.00
θ_6	2.99	3.58	3.46
θ_7	0.17	0.76	0.32
ММР	К-С	К-фон М	А-Д
θ_1	0.85	1.44	0.78
θ_2	0.88	1.45	0.80
θ_3	0.70	1.58	0.70
θ_4	0.13	0.22	0.11
θ_5	9.33	10.11	6.10
θ_6	3.58	8.70	4.56
θ_7	0.78	0.68	0.35

Таблица 4 Среднеквадратичная ошибка оценок параметра $\boldsymbol{\mu}$

	ММП	моменты	L-моменты
θ_1	0.0012	0.0021	0.0017
θ_2	0.0012	0.0022	0.0016
θ_3	0.0009	0.0018	0.0013
θ_4	0.0012	0.0026	0.0017
θ_5	0.0013	0.0032	0.0020
θ_6	0.0071	0.0084	0.0080
θ_7	0.0005	0.0021	0.0009
ММР	К-С	К-фон М	А-Д
θ_1	0.0022	0.0036	0.0020
θ_2	0.0022	0.0036	0.0020
θ_3	0.0018	0.0038	0.0018
θ_4	0.0026	0.0042	0.0021
θ_5	0.0032	0.0035	0.0021
θ_6	0.0084	0.0177	0.0103
θ_7	0.0020	0.0018	0.0010

Таблица 5 Среднеквадратичная ошибка оценок параметра $\boldsymbol{\sigma}$

	ММП	моменты	L-моменты
θ_1	$6.80 \cdot 10^{-5}$	$11.19 \cdot 10^{-5}$	$9.19 \cdot 10^{-5}$
θ_2	$6.43 \cdot 10^{-5}$	$11.12 \cdot 10^{-5}$	$7.96 \cdot 10^{-5}$
θ_3	$5.22 \cdot 10^{-5}$	$9.12 \cdot 10^{-5}$	$6.65 \cdot 10^{-5}$
θ_4	$6.55 \cdot 10^{-5}$	$13.33 \cdot 10^{-5}$	$9.07 \cdot 10^{-5}$
θ_5	$7.06 \cdot 10^{-5}$	$16.73 \cdot 10^{-5}$	$10.47 \cdot 10^{-5}$
θ_6	$6.80 \cdot 10^{-5}$	$8.10 \cdot 10^{-5}$	$7.42 \cdot 10^{-5}$
θ_7	$6.18 \cdot 10^{-5}$	$23.30 \cdot 10^{-5}$	$10.54 \cdot 10^{-5}$
ММР	К-С	К-фон М	А-Д
θ_1	$10.85 \cdot 10^{-5}$	$19.81 \cdot 10^{-5}$	$11.23 \cdot 10^{-5}$
θ_2	$10.84 \cdot 10^{-5}$	$18.51 \cdot 10^{-5}$	$10.50 \cdot 10^{-5}$
θ_3	$9.09 \cdot 10^{-5}$	$19.50 \cdot 10^{-5}$	$9.23 \cdot 10^{-5}$
θ_4	$12.79 \cdot 10^{-5}$	$22.43 \cdot 10^{-5}$	$11.61 \cdot 10^{-5}$
θ_5	$16.18 \cdot 10^{-5}$	$17.52 \cdot 10^{-5}$	$10.50 \cdot 10^{-5}$
θ_6	$7.87 \cdot 10^{-5}$	$15.63 \cdot 10^{-5}$	$9.46 \cdot 10^{-5}$
θ_7	$26.08 \cdot 10^{-5}$	$22.64 \cdot 10^{-5}$	$12.05 \cdot 10^{-5}$

Таблица 6 Среднее время работы методов (в секундах)

	ММП	моменты	L-моменты
θ_1	0.032	0.002	0.002
θ_2	0.032	0.002	0.002
θ_3	0.031	0.002	0.002
θ_4	0.031	0.002	0.002
θ_5	0.035	0.002	0.002
θ_6	0.036	0.002	0.002
θ_7	0.031	0.002	0.002
ММР	К-С	К-фон М	А-Д
θ_1	0.416	0.211	0.200
θ_2	0.375	0.225	0.234
θ_3	0.359	0.229	0.224
θ_4	0.411	0.221	0.206
θ_5	0.392	0.275	0.282
θ_6	0.422	0.269	0.270
θ_7	0.426	0.175	0.175

По результатам тестирования методов оценки параметров на модельных данных можно сделать следующие выводы:

1. Для каждого метода качество оценки относительно других методов в целом одинаково для всех параметров. Нет метода, который давал бы значительно лучшую, чем у другого метода, оценку одного параметра и при этом серьёзно проигрывал по другому параметру. Это значит, что можно провести ранжирование методов, одинаковое для всех параметров.
2. С точки зрения точности оценки самым лучшим можно признать метод максимального правдоподобия. За ним идёт метод L-моментов, далее – метод моментов и метод минимизации расстояния Андерсона-Дарлинга, затем – ММР

Колмогорова–Смирнова и, наконец, ММР Крамера–фон Мизеса.

3. Следует отметить, что, хотя метод моментов и метод L-моментов уступают ММП, они всё же дают оценки очень высокой точности и при этом работают на порядок быстрее ММП. Поскольку для моделирования задержек предлагается использовать смесь 3LN распределений, метод L-моментов может быть использован в качестве промежуточного шага в задаче разделения смеси с целью ускорения работы.

8 Запуск на реальных данных

Мы провели несколько запусков рассмотренных выше методов оценки параметров на реальных данных о задержках в коммуникационной среде суперкомпьютера BlueGene/P. Поскольку для реальных данных на данном этапе работы не представляется возможным ввести объективный численный критерий качества, нашей основной целью было визуальное наблюдение полученных функций плотности. Результат можно видеть на рис. 2. Видно, что рассмотренные методы применимы в условиях реальных данных, и полученные распределения хорошо описывают картину задержек.

9 Заключение

В работе обоснована потребность в построении стохастической модели задержек. На основании анализа смежной предметной области (локальные сети и интернет), а также особенностей, присущих коммуникационным средам, предложена стохастическая модель задержек – смесь 3LN распределений. Поскольку задача параметрического восстановления даже одного компонента смеси оказалась нетривиальной, мы провели обзор существующих методов, а также предложили ранее не применявшийся методы минимизации расстояния. Проведённый нами анализ методов на модельных данных показал, что ММП даёт оценки наибольшей точности, однако метод L-моментов даёт хорошие оценки и при этом работает на порядок быстрее.

В дальнейшем результаты работы предполагается использовать для решения задачи разделения смеси 3LN распределений с целью построения точной и при этом компактной модели задержек для использования в задачах динамического планирования выполнения и диагностики кластера.

Благодарности

Работа выполнена при частичной поддержке РФФИ, проекты 15-07-09214, 16-57-45054, 16-01-00196.

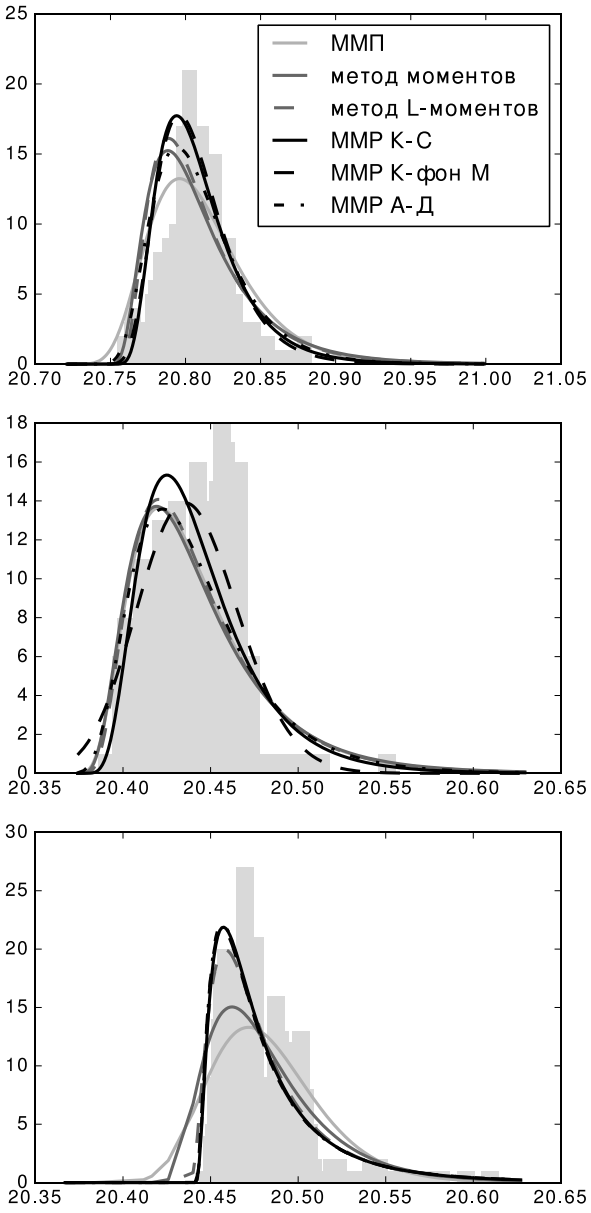


Рисунок 2 Работа методов оценки параметров на реальных данных о задержках для суперкомпьютера BlueGene/P. Линиями показаны восстановленные плотности распределений, светлая столбчатая диаграмма на фоне показывает реальные данные

Литература

[1] Basu, A., Shioya, H., Park, C.: Statistical Inference: the Minimum Distance Approach. CRC Press (2011)

[2] Bílková, D.: Three-parametric Lognormal Distribution and Estimating its Parameters using the Method of L-moments. Reprodukce Lidského Kapitálu (2011)

[3] Calitz, F.: Maximum Likelihood Estimation of the Parameters of the three Parameter Lognormal Distribution – a Reconsideration. Australian J. of Statistics, 15 (3), pp. 185-190 (1973)

[4] Cohen, A., Whitten, B.: Parameter Estimation in Reliability and Life Span Models. Marcel Dekker, New York (1988)

[5] Corlett, A., Pullin, D., Sargood, S.: Statistics of One-way Internet Packet Delays. 53rd IETF (2002)

[6] Gorelov, A., Maysuradze, A. Salnikov, A.: Delay Structure Mining in Computing Cluster. CEUR Workshop Proceedings, 1482. Aachen: M. Jeusfeld c/o Redaktion Sun SITE, Informatik V, RWTH Aachen Germany Germany, pp. 546-551 (2015)

[7] Harter, H., Moore, A.: Local-maximum-likelihood Estimation of the Parameters of Three-parameter Lognormal Populations from Complete and Censored Samples. J. of the American Statistical Association, 61 (315), pp. 842-851 (1966)

[8] Hill, B.: The Three-parameter Lognormal Distribution and Bayesian Analysis of a Point-source Epidemic. J. of the American Statistical Association, 58 (301), pp. 72-84 (1963)

[9] Hosking, J.: L-Moments: Analysis and Estimation of Distributions Using Linear Combinations of Order Statistics. J. of the Royal Statistical Society. Series B (Methodological), 52 (1), pp. 105-124 (1990)

[10] Karakaş, M.: Determination of Network Delay Distribution over the Internet. Citeseer (2003)

[11] Mukherjee, A.: On the Dynamics and Significance of Low Frequency Components of Internet Load. Technical Reports (CIS), 300 p. (1992)

[12] Salnikov, A.: Parus: A Parallel Programming Framework for Heterogeneous Multiprocessor Systems. *Lecture Notes in Computer Science*, 4192, pp. 408-409 (2006)

[13] Salnikov, A., Andreev, D., Lebedev, R.: Toolkit for Analyzing the Communication Environment Characteristics of a Computational Cluster based on MPI Standard Functions. *Moscow University Computational Mathematics and Cybernetics*, 36 (1), pp. 41-49 (2012)

[14] Кобзарь, А.И.: Прикладная математическая статистика. М.: Физматлит (2006)

Обратные задачи моделирования на основе регуляризации и распределенных вычислений в среде Everest

© А.П. Афанасьев © В.В. Волошинов © А.В. Соколов

Институт проблем передачи информации им. А.А. Харкевича РАН,
Москва, Россия

alexander.afanasyev@gmail.com vladimir.voloshinov@gmail.com
alexander.v.sokolov@gmail.com

Аннотация. Изложена методика оценки математических моделей физических явлений, происходящих в некоторой пространственной среде, на основе рядов экспериментальных данных. Целевая функция в обратной оптимизационной задаче идентификации параметров модели включает регуляризирующее слагаемое с неизвестными весовыми коэффициентами при вторых производных функций, описывающими исследуемое явление. Для выбора этих весовых коэффициентов применена процедура перекрестной (взаимной) верификации, когда часть исходных экспериментальных данных используется для «восстановления» остальных. Чем точнее результаты «взаимопроверки» для достаточно широкого набора проверочных тестов, тем «лучше» набор весовых коэффициентов. При выборе направления их улучшения требуется решить большой набор вспомогательных подзадач математического программирования, для чего предложено использовать распределенную систему сервисов оптимизации на платформе Everest.

Ключевые слова: обратные задачи, регуляризация, распределенные вычисления, REST-сервисы, платформа Everest.

Inverse Problem in the Modeling on the Basis of Regularization and Distributed Computing in the Everest Environment

© A.P. Afanasiev © V.V. Voloshinov © A.V. Sokolov

Institute for Information Transmission Problems RAS (Kharkevich institute),
Moscow, Russia

alexander.afanasyev@gmail.com vladimir.voloshinov@gmail.com
alexander.v.sokolov@gmail.com

Abstract. A method for estimating mathematical models of physical spatial phenomena is presented. Estimating is based on the series of experimental data. The objective function in the inverse optimization problem of identification of model parameters includes a regularizing term with unknown weight coefficients for the 2nd derivatives of the spatial function describing the phenomenon. Successive cross-validation procedure is used to choose values of weight coefficients. This cross-validation consists in approximation of one subset of experimental data by processing of a complementary subset. The better accuracy of the "cross-approximation", the better set of weight coefficients. Choosing direction of the possible improvement requires solving a number of subsidiary optimization problems. For that it is proposed to use distributed computing environment of optimization services deployed via Everest toolkit.

Keywords: inverse problems, regularized (ridge) approximation, distributed computing, Everest platform.

Введение

В настоящей работе предложен нестандартный подход к последовательному уточнению математических моделей, описывающих физические явления в некоторой пространственной среде, где для искомых зависимостей имеет смысл понятие «гладкости» по переменным, описывающим состояние модели. Речь может идти о физических процессах, которые моделируются дважды дифференцируемыми функциями на прямой,

Труды XIX Международной конференции «Аналитика и управление данными в областях с интенсивным использованием данных» (DAMDID/ RCDL'2017), Москва, Россия, 10–13 октября 2017 года

плоскости или в трехмерном пространстве.

Данный подход является развитием методов оптимизационной параметрической идентификации на основе экспериментальных данных с учетом возможных (и неизвестных) неточностей в этих данных. Чем шире доступный набор измерений (экспериментальных данных) и выше их точность, тем более точную и подробную модель можно построить. Иногда недостаток количественной информации можно компенсировать дополнительными знаниями о закономерностях исследуемого процесса, уравнений его описывающих и т. д.

Аналогом является обработка статистических данных на основе сплайн-аппроксимации («сглаженных» кубических сплайнов), где коэффициент штрафа за интеграл квадрата 2-й производной аппроксимирующей зависимости играет роль параметра регуляризации [3, 4, 9, 12, 13]. Особенность предлагаемого подхода – минимизация суммы отклонения от экспериментальных данных и штрафа за «негладкость» при дополнительных ограничениях (соотношениях исследуемой модели).

По сути предлагается схема постепенного уточнения математической модели наблюдаемого физического явления с количественной оценкой качества такого уточнения. Если предсказательная точность новой модели повысилась, мы на верном пути.

1 Описание метода

Опишем общую схему метода на примере построения модели, описывающей исследуемое явление с помощью переменных *x, y, z*. Здесь *z* является измеряемой характеристикой, зависящей от *x* и *y*. Требуется построить модель в форме явной и неявной зависимостей между указанными переменными:

$$z = f(x, y),\ x \in X,\ y \in Y; \quad M(x, y, z) = 0. \quad (1)$$

Здесь *f(x,y)* неизвестная функция, которую и требуется «восстановить» при дополнительных предположениях о «физике» наблюдаемого явления, X, Y – множества (интервалы) допустимых значений соответствующих переменных. Предполагаемая, и подлежащая верификации, физическая модель представлена вторым уравнением (1). Неявная зависимость *M* определяет дополнительные связи между переменными. Их может быть несколько, т. е. *M* – вектор-функция. Будем искать функцию *f(x,y)* либо в виде значений на достаточно «мелкой» сетке по переменным *x,y*, либо в классе функций, зависящих от параметров, значения которых и нужно определить, решив обратную задачу.

Пусть у исследователя имеется набор экспериментальных данных (измерений) следующего вида:

$$\{z_k, x_k, y_k\},\ k \in \mathrm{K},\ \mathrm{K} = 1, \ldots k_{\max}, \quad (2)$$

где значения z_k измерены с некоторыми неизвестными погрешностями. Задача восстановления функции *f* часто является некорректной и требует регуляризации. Будем искать зависимость *f(x,y)* (в параметризованном или «сеточном» виде) методом регуляризованной идентификации SvF, Smoothness-vs-Fitting, состоящем в поиске компромисса между точностью совпадения с экспериментальными данными и гладкостью искомой функции *f(x,y)*.

Для формализации такого подхода введем:

- характеристику совпадения с измерениями

$$\Delta_F\left(f(\cdot_{xy}), \mathrm{K}\right) = \frac{1}{|\mathrm{K}|}\sum_{k\in\mathrm{K}}\left(z_k - f(x_k, y_k)\right)^2; \quad (3)$$

- характеристику негладкости

$$\Delta_S\left(f(\cdot_{xy}), \beta_x, \beta_y\right) =$$
$$\beta_x^2 \iint_{X\,Y}\left(\frac{\partial^2 f}{\partial x^2}\right)^2 \mathrm{d}x\mathrm{d}y + 2\beta_x\beta_y \iint_{X\,Y}\left(\frac{\partial^2 f}{\partial x \partial y}\right)^2 \mathrm{d}x\mathrm{d}y + \quad (4)$$
$$\beta_y^2 \iint_{X\,Y}\left(\frac{\partial^2 f}{\partial y^2}\right)^2 \mathrm{d}x\mathrm{d}y$$

(для сеточной функции вторые производные заменяются разностными выражениями);

- компромиссный критерий (функционал Тихонова [10]), зависящий от параметров β_x, β_y и множества измерений, характеризуемого набором индексов K:

$$\Delta\left(f(\cdot_{xy}), \beta_x, \beta_y, \mathrm{K}\right) = \Delta_F\left(f(\cdot_{xy}), \mathrm{K}\right) + \Delta_S\left(f(\cdot_{xy}), \beta_x, \beta_y\right) \quad (5)$$

Будем искать функцию *f(x,y)* в классе дважды непрерывно-дифференцируемых функций, решая следующую задачу минимизации, где переменными являются либо значения функции «на сетке» по переменным *x,y*, либо параметры *f(x,y)*:

$$f^*(\cdot_{xy}) = \underset{f(\cdot_{xy})}{Arg\min}\left\{\begin{array}{l}\Delta\left(f(\cdot_{xy}), \beta_x, \beta_y, \mathrm{K}\right): \\ M\left(x, y, f(x, y)\right) = 0, \\ x \in X, y \in Y.\end{array}\right\} \quad (6)$$

Решение задачи (6) для различных значений β_x, β_y дает функции *f(x,y)*, соответствующие различным соотношениям между «гладкостью» и «точностью» (восстановления экспериментальных данных). Для поиска «разумного» баланса, выраженного в значениях β_x, β_y, воспользуемся процедурой перекрестной верификации [3, 4]. Для этого разобьем множество индексов экспериментальных данных на набор непересекающихся подмножеств

$$\mathrm{K} = \bigcup_{i\in I}\mathrm{K}_i,\ \mathrm{K}_i \cap \mathrm{K}_j = \varnothing,\ i \neq j. \quad (7)$$

Уберем из множества K одно из подмножеств K_i. Решим задачу минимизации (6) на оставшемся наборе данных $\mathrm{K}\setminus\mathrm{K}_i$. Пусть ее решение $f^*_{\mathrm{K}_i}(x, y)$:

$$f^*_{\mathrm{K}_i}(\cdot_{xy}) = \underset{f(\cdot_{xy})}{Arg\min}\left\{\begin{array}{l}\Delta\left(f(\cdot_{xy}), \beta_x, \beta_y, \mathrm{K}\setminus\mathrm{K}_i\right): \\ M\left(x, y, f(x, y)\right) = 0, \\ x \in X, y \in Y.\end{array}\right\} \quad (8)$$

Определим отклонение $f^*_{\mathrm{K}_i}(x,y)$ от измерений для $k \in \mathrm{K}_i$ по формуле $\sum_{k\in \mathrm{K}_i}\left(z_k - f^*_{\mathrm{K}_i}(x_k,y_k)\right)^2$.

Повторив эту процедуру для всех подмножеств K_i, $i \in I$, получим «перекрестную» оценку точности модели для заданных β_x, β_y:

$$\Phi(\beta_x,\beta_y) = \left(\sigma_{\mathrm{SvF}}(\beta_x,\beta_y)\right)^2 = \frac{1}{|\mathrm{K}|}\sum_{i\in \mathrm{I}}\sum_{k\in \mathrm{K}_i}\left(z_k - f^*_{\mathrm{K}_i}(x_k,y_k)\right)^2. \quad (9)$$

Для получения минимальной погрешности моделировании и выбора «оптимального» соотношения близость–сложность будем искать β_x, β_y, которые минимизируют величину $\sigma_{\mathrm{SvF}}(\beta_x,\beta_y)$. Для найденных в результате минимизации значений β_x, β_y искомая функция $f^*(x,y)$ определяется в результате решения основной задачи (6). Итоговая погрешность аппроксимации (невязка) определяется по формуле

$$\left(\sigma^*\right)^2 = \frac{1}{|\mathrm{K}|}\sum_{k\in \mathrm{K}}\left(z_k - f^*(x_k,y_k)\right)^2. \quad (10)$$

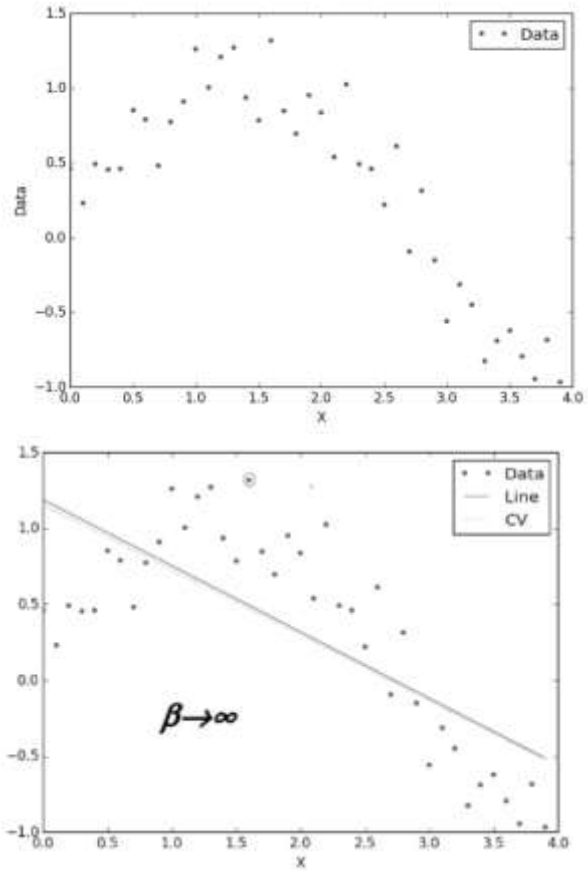


Рисунок 1a Исходные данные и результат метода наименьших квадратов $\beta \to \infty$

Таким образом, процедура расчетов соответствует двухуровневой задаче оптимизации:

$$\Phi(\beta_x,\beta_y) = \frac{1}{|\mathrm{K}|}\sum_{i\in \mathrm{I}}\sum_{k\in \mathrm{K}_i}\left(z_k - f^*_{\mathrm{K}_i}(x_k,y_k)\right)^2 \to \min_{\beta_x,\beta_y \geq 0}, \quad (11)$$

где функции $f^*_{\mathrm{K}_i}(x,y)$ определяются как решения независимых задач оптимизации (8).

Результат метода SvF находится, в некотором смысле, «между» результатами применения хорошо известных методов наименьших квадратов и кубической сплайн-интерполяции. На Рис. 1 метод SvF продемонстрирован на примере восстановления функции одного переменного по набору исходных данных (Рис. 1a) при отсутствии дополнительных модельных соотношений (формально можно положить $\boldsymbol{M}(x,y,z)$ тождественно равной нулю). Для функции одной переменной x требуется лишь один скалярный параметр β. При $\beta \to 0$ задача (6) становится задачей сплайн-интерполяции, решением которой является т. н. кубический сплайн, т. е. функция, имеющая минимальный (см. (4)) интеграл квадрата 2-ой производной и проходящая через все «экспериментальные» точки (2), Рис. 1b).

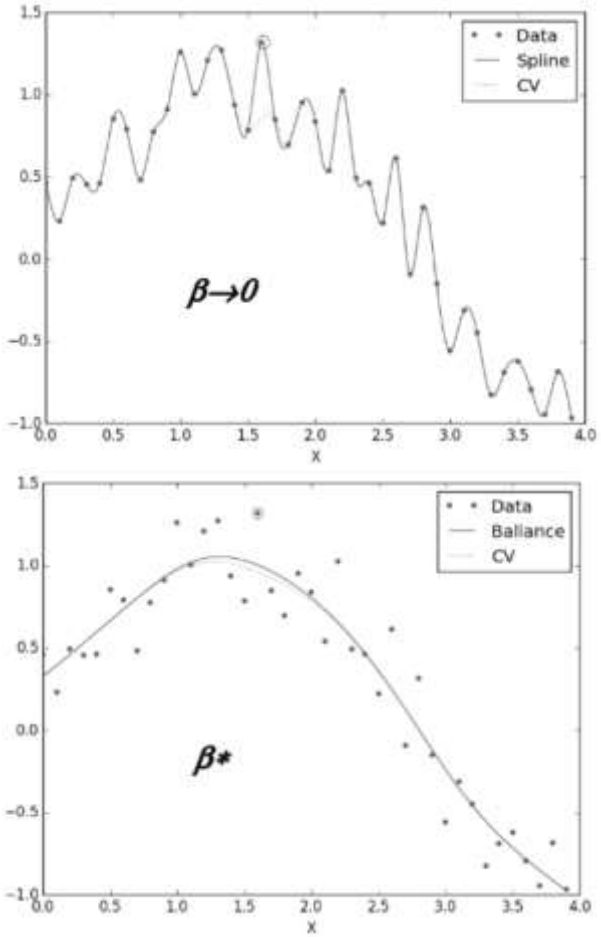


Рисунок 1b Кубический сплайн, $\beta \to 0$, и значение β^*, полученное методом SvF

При $\beta \to \infty$ мы получим линейную функцию, минимизирующую сумму квадратов отклонения от исходных данных, Рис. 1a. «Компромиссный» результат сплайн-аппроксимации со штрафом (за негладкость), определенным методом SvF, представлен на Рис. 1b снизу.

Опишем процедуру решения задачи верхнего уровня (11). Введем обозначение $P(\boldsymbol{a},\boldsymbol{\beta})$ для произвольного многочлена 2-го порядка вектора

переменных $\boldsymbol{\beta}$, зависящего от вектора коэффициентов $\boldsymbol{a}$

$$P(\mathbf{a},\boldsymbol{\beta}) = a_{xx}\beta_x^2 + a_{xy}\beta_x\beta_y + a_{yy}\beta_y^2 + a_x\beta_x + a_y\beta_y + a_0, \quad (12)$$

где $\mathbf{a} = (a_{xx}, a_{xy}, a_{yy}, a_x, a_0), \boldsymbol{\beta} = (\beta_x, \beta_y)$.

Далее алгоритм строит последовательность значений $\boldsymbol{\beta}^{\nu}$, ν=1,2,… . Пусть, на *N*-ом шаге построены значения $\boldsymbol{\beta}^{\nu}$, ν=1:N, для которых вычислены значения $\Phi^{\nu} = \Phi(\boldsymbol{\beta}^{\nu})$. Без ограничения общности (возможно, после перенумерации) можно считать, что Φ^N – минимальное из полученных значений. Будем трактовать $\{\Phi^{\nu}, \boldsymbol{\beta}^{\nu}\}_{\nu=1}^{N}$ как набор точек в R^3. Рассмотрим задачу аппроксимации этих точек графиком многочлена 2-го порядка (12), причем чем вектор $\boldsymbol{\beta}^{\nu}$ ближе к «наилучшему» $\boldsymbol{\beta}^N$, тем с большим весом он будет учитываться. Кроме того, будем штрафовать за кривизну построенной функции с коэффициентом μ (подлежащим выбору):

$$\sum_{\nu=1:N} e^{-\|\boldsymbol{\beta}^{\nu}-\boldsymbol{\beta}^N\|}\left(\Phi^{\nu} - P(\mathbf{a},\boldsymbol{\beta}^{\nu})\right)^2 + \mu\left(a_{xx}^2 + a_{xy}^2 + a_{yy}^2\right) \to \min_{\mathbf{a}}. \quad (13)$$

Пусть $\mathbf{a}^*(\mu)$ оптимальное значение вектора переменных в этой задаче выпуклого программирования.

Выберем значение μ, минимизируя отклонение значения аппроксимирующего полинома от наилучшего значения Φ^N. Тем самым, для аппроксимации получаем вновь двухуровневую задачу:

$$\left|P(\mathbf{a}^*(\mu),\boldsymbol{\beta}^N) - \Phi^N\right| \to \min_{\mu\geq 0}, \quad (14)$$

где $\mathbf{a}^*(\mu)$ – решение задачи «нижнего уровня» (13).

Здесь в задаче верхнего уровня нужно выбрать единственную переменную μ, а задача нижнего уровня эффективно разрешима благодаря ее выпуклости. Поэтому для поиска оптимального штрафного коэффициента μ^* применим тот или иной алгоритм минимизации функции одного переменного.

После аппроксимации зависимости $\Phi(\boldsymbol{\beta})$ квадратичной функцией $P(\mathbf{a}^*(\mu^*),\boldsymbol{\beta})$, новое значение вектора $\boldsymbol{\beta}$ находится в результате решения следующей вспомогательной задачи, которая напоминает метод линеаризации Пшеничного–Данилина [11], применяемый к минимизации функции $P(\mathbf{a}^*(\mu^*),\boldsymbol{\beta})$ по $\boldsymbol{\beta}$:

$$\left(\nabla_{\boldsymbol{\beta}} P(\mathbf{a}^*(\mu^*),\boldsymbol{\beta}^N)\right)^{\mathrm{T}}(\boldsymbol{\beta}-\boldsymbol{\beta}^N) + \frac{1}{2}\left\|\boldsymbol{\beta}-\boldsymbol{\beta}^N\right\|^2 \to \min_{\beta\geq 0},$$

где $\nabla_{\boldsymbol{\beta}} P(\ldots)$ обозначает градиент многочлена (12) по переменным β_x, β_y. Заметим, что решение этой выпуклой задачи квадратичного программирования существует и единственно. Вычислим значение $\Phi^{N+1} = \Phi(\boldsymbol{\beta}^{N+1})$, решив набор задач нижнего уровня (8). Если величина $\left|\Phi(\boldsymbol{\beta}^{N+1}) - \Phi(\boldsymbol{\beta}^N)\right|$ меньше некоторого порогового значения, работа алгоритма прекращается. Если это не так, то схема расчетов повторяется для расширенного набора $\{\Phi^{\nu}, \boldsymbol{\beta}^{\nu}\}_{\nu=1}^{N+1}$.

3 Демонстрация метода при моделировании распространения тепла в высокотемпературной плазме

Физические явления в горячей плазме, удерживаемой сильным магнитным полем в тороидальных вакуумных камерах установок термоядерного синтеза, таких, как ТОКАМАКи и стеллараторы, важны для перспектив термоядерной энергетики. Например, в экспериментах наблюдается «быстрый нелокальный перенос тепла» – практически мгновенное (по сравнению с «классической» тепловой диффузией) повышение температуры в центре плазменного шнура после охлаждения его периферии или обратный процесс – мгновенное понижение температуры в центре при быстром нагреве периферии плазмы.

Здесь мы не будем обсуждать неясную физику этого явления. Ограничимся демонстрацией применения метода к предварительной обработке экспериментальных данных из статьи [8], где обсуждается явление «быстрого» охлаждения плазмы в стеллараторе LHD, http://www.lhd.nifs.ac.jp/en

Исходными данными являются графики зависимостей температуры плазмы от расстояния (ρ) до центра тороидальной камеры в различные моменты времени (t), Рис. 2.

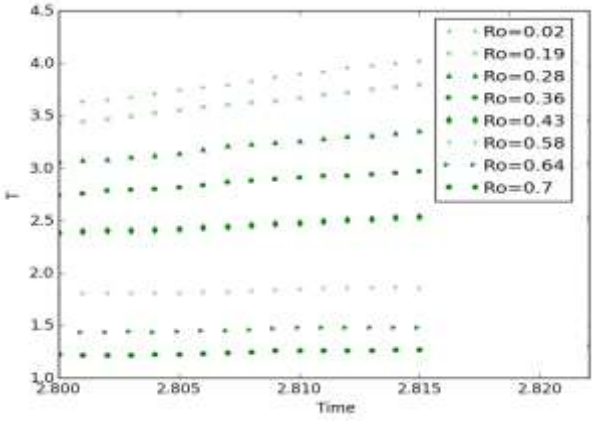


Рисунок 2 Температура на разных расстояниях от центра тороидальной камеры $T(\rho,t)$, [8]

Множеством $\mathbb{K}$ экспериментальных данных о температуре, см. (2), является набор пар индексов i_ρ, i_t, значений расстояния и моментов времени. Разделим множество измерений температуры на 8 частей, определяемых значениями ρ (горизонтальные группы точек на Рис. 2). Для перекрестной верификации будем использовать 6 множеств (см. (7)):

$$\mathrm{K_i} = \{(i_\rho, i_t): i_t \in I_t\}, \; i_\rho = 2:7. \quad (15)$$

Крайние значения (ρ_1=0.02 и ρ_9=1) в перекрестном оценивании не учитывались, т. к. в этой задаче важна интерполяция значений температуры плазмы.

Ниже приведен ряд моделей наблюдаемого явления вида (1). Они отличаются друг от друга предположениями о процессе распространения тепла в плазме в форме второй группы соотношений (1). Цель – выбрать модель, которая бы максимально точно описывала функцию $T(\rho,t)$ – зависимость температуры плазменного пучка от ρ и t. Можно ожидать, что при «правильных» уточнениях *предсказательная точность* моделирования (по отклонению функции $T(\rho,t)$ от значений на Рис. 2) должна улучшиться. Результаты расчетов по всем моделям приведены ниже в Таблице 1. Графики, иллюстрирующие расчеты по моделям, доступны на странице http://distcomp.ru/~vladimirv/damdid2017.

3.1 Модель 1 (сплайн-интерполяция без SvF)

Пусть у нас нет никаких предположений о физических причинах наблюдаемого явления. Будем искать функцию $T(\rho,t)$, проходящую через все точки используемого набора измерений и имеющую минимальную кривизну в смысле формулы (4), а точность определять процедурой перекрестной верификации на подмножествах (15). Формально этот прием соответствует случаю, $\beta\to 0$, Рис. 1b.

Получили задачу сплайн-интерполяции. При достаточно «необременительных» предположениях о расположении узлов интерполяции её решение существует и единственно [13].

3.2 Модель 2 (Сплайн-аппроксимация)

Не имея, как и выше, никаких предположений о «физике» явления, применим метод SvF, когда функция $M(x,y,z)$ тождественно равна нулю (см. (1) и Рис. 1b, снизу). Эта задача является задачей сплайн-аппроксимации с выбором штрафа за негладкость на основе перекрестного оценивания. Как и для Модели 1, её решение существует и единственно [13]. Несмотря на то, что построенная функция температуры уже не проходит через узлы, оценка точности заметно улучшилась (см. Табл. 1).

3.3 Модель 3 (Метод SvF и простое дифференциальное уравнение)

Предположим, что процесс описывается простейшим дифференциальным уравнением:

$$\frac{\partial T}{\partial t}(\rho,t)=\pi(\rho,t) \tag{16}$$

Здесь неизвестными являются две функции $T(\rho,t)$ и $\pi(\rho,t)$ от двух переменных. Поэтому в схеме **SvF** проводится оптимизация по четырем коэффициентам штрафа «за негладкость» обеих функций: $\beta_{T\rho}$, β_{Tt} и $\beta_{\pi\rho}$, $\beta_{\pi,t}$. Поскольку для неизвестной функции $\pi(\rho,t)$, правой части дифференциального уравнения (16), нет экспериментальных данных, то основной компромиссный «критерий» (5) имеет вид

$$\begin{aligned}\Delta\left(\left\{T(\cdot_{\rho t}),\pi(\cdot_{\rho t})\right\},\beta_{T\rho},\beta_{Tt},\beta_{\pi\rho},\beta_{\pi t},\mathrm{K}\right)=\\ \Delta_F\left(T(\cdot_{\rho t}),\mathrm{K}\right)+\Delta_S\left(T(\cdot_{\rho t}),\beta_{T\rho},\beta_{Tt}\right)+\\ \Delta_S\left(\pi(\cdot_{\rho t}),\beta_{\pi\rho},\beta_{\pi t}\right)\end{aligned} \tag{17}$$

Сравнение с предыдущей моделью (см. Табл. 1) не выявило принципиальных изменений.

3.4 Модель 4 (Метод SvF и тепловая диффузия)

Здесь предполагается, что распространение тепла описывается классическим дифференциальным уравнением тепловой диффузии в цилиндрически симметричной среде (как приближении тороидальной камеры) с заранее неизвестным коэффициентом «теплопроводности» χ.

$$\frac{\partial T}{\partial t}(\rho,t)=\chi\frac{\partial}{\rho\partial\rho}\left(\rho\frac{\partial}{\partial\rho}\left(T(\rho,t)-T_0(\rho)\right)\right), \tag{18}$$

где $T_0(\rho)$ – функция температуры пучка в начальный момент времени предполагается известной. В такой постановке критерий (5) имеет вид (17), но, кроме коэффициентов $\beta_{T\rho}$, β_{Tt}, $\beta_{\pi\rho}$ и $\beta_{\pi t}$, минимизация проводится еще по переменной χ. Результаты моделирования показывают, что учет только диффузионного члена (без дополнительного «источника») не является удачным: точность моделирования падает.

3.5 Модель 5 (Метод SvF, тепловая диффузия и «неизвестный источник»)

Предположим, что наряду с «медленной» тепловой диффузией в плазме действует некоторый дополнительный механизм переноса тепла, который в следующей формуле обозначен $S(\rho,t)$:

$$\frac{\partial T}{\partial t}(\rho,t)=\chi\frac{\partial}{\rho\partial\rho}\left(\rho\frac{\partial}{\partial\rho}\left(T(\rho,t)-T_0(\rho)\right)\right)+S(\rho,t). \tag{19}$$

Здесь функция критерия (5) имеет вид, аналогичный (17), но зависит от четырех β-коэффициентов для функций $T(\rho,t)$ и $S(\rho,t)$:

$$\begin{aligned}\Delta\left(\left\{T(\cdot_{\rho t}),S(\cdot_{\rho t})\right\},\beta_{T\rho},\beta_{Tt},\beta_{S\rho},\beta_{St},\mathrm{K}\right)=\\ =\Delta_F\left(T(\cdot_{\rho t}),\mathrm{K}\right)+\Delta_S\left(T(\cdot_{\rho t}),\beta_{T\rho},\beta_{Tt}\right)+\\ +\Delta_S\left(S(\cdot_{\rho t}),\beta_{S\rho},\beta_{St}\right).\end{aligned} \tag{20}$$

В итоге определяются коэффициенты $\beta_{T,\rho}$, $\beta_{T,t}$, $\beta_{S,\rho}$ и $\beta_{S,t}$, и коэффициент «теплопроводности» χ. Точность моделирования заметно возросла (Табл. 1).

3.6 Сравнение результатов расчетов

Как уже было объявлено, результаты расчетов сведены в Таблицу 1. Графические изображения построенных зависимостей от ρ и t приведены на рисунках в http://distcomp.ru/~vladimirv/damdid2017.

Заметим, что переход от «простых» сплайнов к более содержательным моделям 2–4 дал незначительное улучшение точности перекрестной проверки. Но расчет по Модели 5 дал многократное улучшение показателей точности, т. е. выделение неизвестного «источника» (переносчика тепловой

энергии) является, по-видимому, верным уточнением модели. Приведенный пример демонстрирует применение метода для количественной проверки «качества» различных математических моделей на имеющемся наборе экспериментальных данных.

Таблица 1 Результаты моделирования

Модель	Погрешность	
	Кросс-валидации, %	Аппроксимации, СКО[3], %
1[4]	11.94	0
2	9.25	3.55
3	9.19	3.55
4	10.00	8.53
5 (χ=0.21)	1.85	0.59

4 Возможности программной реализации в среде Everest

Предлагаемая методика основана на решении задач математического программирования. Для ее практического применения нужен набор программных инструментов для: 1) описания указанных задач; 2) формирования структур данных, соответствующих отдельным экземплярам таких задач; 3) отправки этих данных пакетам численных методов (решателям), для поиска решения; 4) обработки результатов работы решателей, например, для изменении метода расчетов. Сложившаяся практика применения оптимизационных моделей предусматривает два способа организации расчетов.

Первый, «низкоуровневый», на основе открытого программного интерфейса (API) решателя для некоторого языка программирования (C/C++, C#, Java, Python и т. п.). Подготовка данных для отправки решателю и обработка результатов производятся обычно на языке API решателя. Такой подход, хотя и может привести к созданию высокопроизводительной системы расчетов, является достаточно трудоемким и требует привлечения высококвалифицированных программистов. Кроме того, изменения в схеме расчетов или структуре применяемой оптимизационной модели требуют переписывания значительных фрагментов программного кода.

Второй, «высокоуровневый», подход использует алгебраические (декларативные) языки оптимизационного моделирования, что гораздо удобнее, особенно для поисковых исследований. Развитие таких языков (AML, Algebraic Modelling Language – в англоязычной литературе) ведется уже более 30 лет. До настоящего времени наиболее популярными являются AMPL, GAMS. Основными составляющими AML-систем являются: собственно язык для описания оптимизационных моделей, средства автоматического дифференцирования и унифицированный интерфейс взаимодействия с пакетами.

AML-языки позволяют записать соотношения оптимизационных задач (параметры, переменные, целевую функцию, ограничения, индексы параметров, переменных и ограничений и т. п.), разделив «символьное описание» задачи (т. н. *модельное представление*) и «конкретные данные» (наборы индексов, значения числовых параметров и т. п.). Символьная модель и конкретные данные, представленные обычно текстовыми файлами, обрабатываются специальным транслятором. На выходе – специальная структура данных в виде т. н. стаб-файла (stub, в терминологии AMPL), готового для передачи решателям, совместимым с языком моделирования. Для нелинейных задач стаб также содержит правила вычисления первых и вторых производных всех функций задачи математического программирования. Если численный метод находит решение, то AML-совместимый решатель возвращает файл, содержащий значения всех «прямых» и двойственных переменных задачи (множителей Лагранжа при ограничениях). Формат этого файла соответствуют стандарту применяемого AML, и AML-транслятор может его импортировать.

Также эти языки позволяют описывать сложносоставные сценарии расчетов по оптимизационным моделям: содержащие условные переходы, циклы, динамическое формирование новых задач на основе результатов решения предыдущих, создание наборов задач для различных наборов значений параметров и т. п. Являясь по назначению языками программирования высокого уровня, AMPL и GAMS не отвечают требованиям, предъявляемым даже к процедурным языкам (надо иметь ввиду «почтенный возраст» языков, AMPL и GAMS появились в конце 1970-х годов). Например, в них нет понятия процедуры-функции, все переменные (кроме внутренних индексов циклов или операторов «итерирования») являются глобальными и т. п.

В связи с этим большой интерес вызывает система оптимизационного моделирования Pyomo (PYthon Optimization Modeling Objects) [2] http://pyomo.org, основанная на популярном объектно-ориентированном языке программирования Python (Pyomo представляет собой специализированный Python-пакет). Четыре года лет назад Pyomo стал совместимым со стандартом AMPL. Это произошло благодаря тому, что авторы AMPL 12 лет назад «раскрыли» внутренний формат AMPL-стаба.

Принцип расчетов в системе Pyomo повторяет схему применения языка AMPL: модель (в форме набора Python-объектов) вместе с исходными данными (либо в виде Python-объектов, либо в формате файлов с данными AMPL-формате) преобразуются в AMPL-стаб, передаваемый AMPL-решателю. Файл с решением можно считать специальной процедурой пакета Pyomo, для оформления результата и/или подготовки исходных данных новых задач математического программирования.

[3] Среднеквадратичное отклонение

[4] Сплайн-интерполяция

4.1 Сведения о платформе Everest

Разработка программного обеспечения для создания систем на основе REST-сервисов ведутся в нашем коллективе около шести лет. Первоначальная и последующая стабильная версии ПО имели название MathCloud, mathcloud.org. Последние три года велась активная работа по переходу на новую версию программного инструментария, т. н. Everest [6, 7], http://everest.distcomp.org.

Программный инструментарий Everest является системой с открытым кодом, свободно доступным на популярном портале gitlab.com. Семантика Everest базируется на следующей иерархии понятий:

- *Приложение Everest* – REST-сервис с REST-интерфейсом в формате JSON; приложение Everest, вообще говоря, является абстракцией, для которой не выделено никакого реального вычислительного ресурса (его выбор и подключение происходят непосредственно перед вызовом);
- *Вычислительный ресурс Everest* – реальное вычислительное устройство или инфраструктура (сервер, кластер, грид, облако), где производится обработка данных Everest-приложениями;
- *Агент доступа к вычислительным ресурсам Everest* – программный модуль (на Python), обеспечивающий подключение ресурса к системе Everest (некоторые основные типы ресурсов представлены на Рис. 3);
- *сервер Everest* (контейнер приложений) – центральный сервер для: сохранения дескрипторов всех приложений; регистрации пользователей; управления правами доступа к созданным приложениям; управление очередями заданий;
- веб-интерфейс работы с сервером Everest, включающий средства создания приложений, запуска и контроля за ходом выполнения заданий.

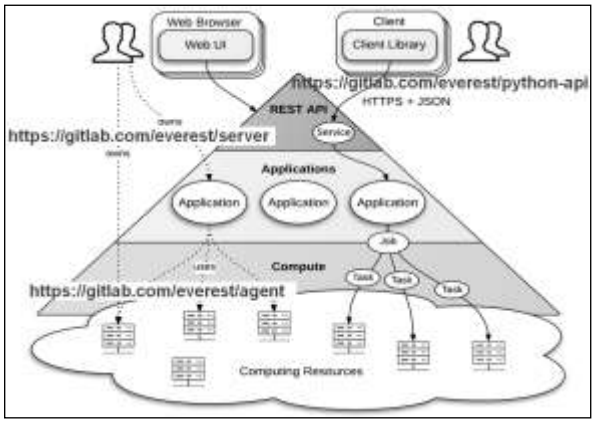


Рисунок 3 Архитектура программного инструментария Everest

Перечислим ряд особенностей Everest, важных с точки зрения практического развертывания и применения систем оптимизационного моделирования в распределенной вычислительной инфраструктуре.

1. Автор приложения может «незаметно» для пользователей повышать/понижать вычислительную «мощность» приложений (сервисов), изменив список ресурсов (фактически агентов доступа к ресурсам), ассоциированных с данным приложением. Например, если комплект решателей будет установлен на новом вычислительном сервере вместе с агентом Everest, то производительность Everest-приложения для решения задач оптимизации повысится.

2. Платформа Everest предлагает унифицированный программный интерфейс (Everest Python API), gitlab.com/everest/python-api. Он позволяет клиентским модулям на Python взаимодействовать с сервисами Everest по модели асинхронных вызовов удаленных объектов и программировать вычислительные сценарии координированной обработки данных несколькими приложениями Everest. При этом независимые задания будут выполняться одновременно несколькими приложениями (или одним приложением, но на разных вычислительных ресурсах, подключенных к этому приложению). Задача балансировки вычислительной нагрузки между ресурсами возложена на сервер Everest.

3. Подсистема балансировки вычислительной нагрузки управляет выполнением заданий, распределяя их между вычислительными ресурсами, подключенными к одному приложению. Пользователь может не знать, какие именно ресурсы обрабатывают его данные. Эта подсистема постоянно совершенствуется разработчиками Everest, в частности, ожидается возможность выбора различных политик распределения заданий между ресурсами.

4. Система контроля доступа к приложениям и защиты данных в Everest использует две технологии: защищенный обмен данными между Everest-сервером и агентами по протоколу HTTPS/SSH; специальные «временные» ключи, т. н. токены, выдаваемые зарегистрированным пользователям Everest с ограниченным сроком действия (7 дней). Предъявление токенов обязательно и для работы с веб-интерфейсом сервера, и при вызове приложений через Everest API.

4.2 Сервис оптимизации

Базовым сервисом решения задач оптимизации в Everest является сервис solve-ampl-stub решения задач математического программирования, представленных в виде AMPL-стаб-файла [1], пакетом численных методов (решателем), указанным при обращении к сервису. Сейчас сервис обеспечивает унифицированный доступ к следующим пакетам, позволяющим решать основные типы задач математического программирования (LP/MILP/NLP/MINLP):

- **Ipopt** (Coin-OR **I**nterior **P**oint **Opt**imizer, NLP), https://projects.coin-or.org/Ipopt;
- **CBC** (**C**oin-OR **B**ranch-and-**C**ut, LP, MILP), https://projects.coin-or.org/Cbc;

- **SCIP** (**S**olving **C**onstraint **I**nteger **P**rograms, LP, MILP, MINLP (билинейные невыпуклые)), http://scip.zib.de
- **Bonmin** (COIN-OR **B**asic **O**pen-source **N**onlinear (convex) **M**ixed **In**teger programming, MINLP), https://projects.coin-or.org/Bonmin

Данным приложением можно пользоваться как через его веб-интерфейс, так и через программный интерфейс Everest Python API.

4.3 Сведения о программной архитектуре Pyomo-Everest

Основным требованием к системе было: обеспечить возможность выполнения любых программ (сценариев расчетов) на языке Python с использованием пакета Pyomo и AMPL-совместимых решателей в среде сервисов оптимизации Everest, возможно, после некоторой модификации самой программы согласно определенному набору правил. Общий принцип работы PyomoEverest (https://github.com/ distcomp/pyomo-everest) аналогичен разработанной нами ранее системе AMPLx [5].

PyomoEverest состоит из двух элементов: 1) модуля на Python, который посредством Everest Python API обеспечивает взаимодействие с сервисом solve-ampl-stub (см. выше); 2) пула вычислительных ресурсов, подключенных к сервису solve-ampl-stub посредством агентов доступа Everest.

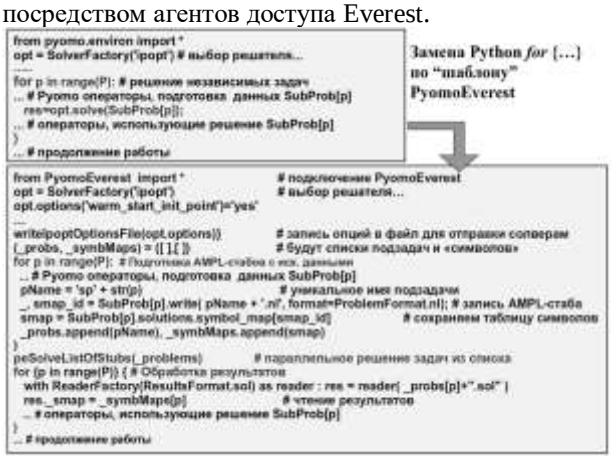


Рисунок 4 Модификация Python/Pyomo кода по «шаблону» PyomoEverest

Типовой прием «распараллеливания» фрагмента алгоритма расчетов, записанного на Pyomo, представлен на Рис. 4. Вверху, в рамке, приведены фрагменты кода для выбора решателя и цикла `for`, где решается набор независимых подзадач. Внизу находится фрагмент модифицированного кода. Правило модификации в том, чтобы *каждый цикл* `for` или `while` нужно заменить тремя группами операторов:

1. цикл формирования набора подзадач в форме AMPL-стабов;
2. параллельное решение задач, представленных своими стаб-файлами, приложением Everest (сейчас – solve-ampl-stub) с подключенным к нему пулом AMPL-совместимых решателей;
3. цикл обработки результатов, доставленных в виде набора файлов *.sol с решениями подзадач.

Модифицированный фрагмент приведен в нижней рамке. Здесь важно использование Python-класса pyomo.core.base.SymbolMap. Экземпляр этого класса (структура symbol_map) создается при записи стаб-файла подзадачи в первом цикле, сохраняется (здесь – в массиве _symbMaps) и применяется во втором цикле при чтении решений из *.sol-файлов для корректного сопоставления их содержимого структуре соответствующей оптимизационной подзадачи.

Заключение

Приведенные результаты моделирования динамики температуры плазмы подтверждают эффективность предложенного метода «гладкой» регуляризации для обработки экспериментальных данных.

Практическое применение метода основано на «перекрестной» (взаимной) верификации – процедуре определения «пропущенной» части экспериментальных данных по остальным измерениям. Это известный байесовский подход к «настройке» параметров некоторой регрессии по статистическим данным [3, 4]. Поскольку здесь требуется решать наборы независимых задач математического программирования, то работа алгоритма может быть ускорена за счет одновременного решения указанных задач пулом решателей, установленных в распределенной вычислительной среде.

Эта вычислительная схема характеризуется периодическим решением относительно небольшого набора (десятки) независимых и относительно сложных задач математического программирования (время решения каждой современным решателем на современном сервере – не менее пары минут). Для ее реализации в режиме распределенных вычислений предлагается использовать систему Pyomo-Everest. Решение всех задач выполняется сервисом оптимизации Everest, причем независимые подзадачи могут решаться параллельно под управлением службы балансировки вычислительной нагрузки Everest на ресурсы, подключенных к сервису.

Предложенный подход указывает перспективное направление применения распределенных систем на основе высокоуровневых средств оптимизационного моделирования.

Финансовая поддержка

Работа поддержана Российским научным фондом (грант № 16-11-10352).

Литература

[1] Fourer, R., Gay, D.M., Kernighan, B.W.: AMPL: A Modeling Language for Mathematical Programming, 2nd edition.: Duxbury Press (2002)

[2] Hart, W.E., Laird, C., Watson, J.-P., Woodruff, D.L.: Pyomo-optimization modeling in python, 67, 238 p. Springer (2012)

[3] Hastie, T.J., Tibshirani, R.J.: Generalized additive models, 43, CRC Press (1990)

[4] Hastie, T.J., Tibshirani, R.J., Friedman, J.: Unsupervised Learning. The Elements of Statistical Learning: Springer, pp. 485-585 (2009)

[5] Smirnov, S., Voloshinov, V., Sukhosroslov, O.: Distributed Optimization on the Base of AMPL Modeling Language and Everest Platform. Procedia Computer Science, 101, pp. 313-322 (2016)

[6] Sukhoroslov, O., Rubtsov, A., Volkov, S.: Development of Distributed Computing Applications and Services with Everest Cloud Platform. Computer Research and Modeling, 7 (3), pp. 593-599 (2015)

[7] Sukhoroslov, O., Volkov, S., Afanasiev, A.A.: Web-Based Platform for Publication and Distributed Execution of Computing Applications. Parallel and Distributed Computing, 14th Int. Symposium on IEEE, pp. 175-184 (2015)

[8] Tamura, N., et al.: Impact of Nonlocal Electron Heat Transport on the High Temperature Plasmas of LHD. Nuclear Fusion, 47 (5), pp. 449 (2007)

[9] Линник, В.Г., Соколов, А.В., Мироненко, И.В.: Паттерны 137CS и их трансформация в ландшафтах ополья Брянской области. Современные тенденции развития биогеохимии. М.: ГЕОХИ РАН, сс. 423-434 (2016)

[10] Морозов, В.А.: Регулярные методы решения некорректно поставленных задач. М.: Наука (1987)

[11] Пшеничный, Б.Н.: Метод линеаризации. М.: Наука (1983)

[12] Роженко, А.И.: Теория и алгоритмы вариационной сплайн-аппроксимации. Новосибирск: Изд-во ИВМиМГ СО РАН (2005)

[13] Тихонов, А.И.: О математических методах автоматизации обработки наблюдений. Проблемы вычислительной математики. М.: МГУ, сс. 3-17 (1980)

Development of Data-Intensive Services with Everest

© Oleg Sukhoroslov © Alexander Afanasiev

Institute for Information Transmission Problems of the Russian Academy of Sciences, Moscow, Russia

sukhoroslov@iitp.ru afanasiev@iitp.ru

Abstract. The paper considers development of domain-specific web services for processing of large volumes of data on high-performance computing resources. The development of these services is associated with a number of challenges, such as integration with external data repositories, implementation of efficient data transfer, management of user data stored on the resource, execution of data processing jobs and provision of remote access to the data. An approach for building big data processing services on the base of Everest platform is presented. The proposed approach takes into account the characteristic features and supports rapid deployment of these services on the base of existing computing infrastructure. An example of service for short-read sequence alignment that processes the next-generation sequencing data on a Hadoop cluster is described.

Keywords: big data, web services, data transfer, data management, distributed data processing.

1 Introduction

The explosive growth of data, observed in a variety of areas from research to commerce, requires the use of high-performance resources and efficient means for storing and processing large amounts of data. During the last decade, the distributed data processing technologies like Hadoop and Spark are emerged. However, the complexity of the hardware and software infrastructure prevents its direct use by non-specialists, and requires the creation of user-friendly tools to solve particular classes of problems.

One way of implementing such tools is the creation of domain-specific services based on the Software as a Service model. This model allows users of such services to quickly, without installing software, reuse ready-made implementations of data processing methods in a particular domain. At the same time, the user does not need to delve into the peculiarities of storing and processing data on high-performance resources behind these services.

Data-intensive services (DIS), in comparison to conventional computational services with a small amount of data, started to develop recently, so the principles and variants of implementation of these services are poorly understood. There are several academic projects aimed at supporting specific areas of research, for example, the Globus Genomics [1] service for analyzing the next-generation sequencing data and the PDACS portal [2] for storing and analyzing data in the cosmology domain. The first system uses Amazon cloud resources as a computing infrastructure, while the second uses the resources of the NERSC and Magellan science cloud. Commercial cloud solutions, such as Amazon ML, Microsoft Azure ML, Databricks Cloud, are general-purpose platforms, including a set of universal services, as well as its own infrastructure for storing and processing data.

There is a lack of best practices for implementation of DIS on the basis of the existing infrastructure for big data processing such as a cluster running Hadoop or Spark platforms which are increasingly used for the analysis of scientific data [3-5]. Also, little attention is paid to the integration of DIS with existing repositories and data warehouses, including the cloud-based ones, as well as other services. A lot of experience in the integration of distributed resources for storing and processing data has been accumulated within the grid infrastructures [6], however these environments are complex for use by researchers and do not support the use of new models of computations and technologies such as Hadoop. Finally, there is a lack of platforms for implementation and deployment of DIS that would provide ready-made solutions of typical problems encountered when creating this kind of services.

This work is designed to fill these gaps. Chapter 2 describes the characteristics and requirements for DIS. Chapter 3 discusses the principles of implementation of the DIS based on the Everest platform, initially focused on creating services working with a small amount of data. A distinctive feature of the proposed approach is support for the rapid implementation of DIS based on available computing resources and data warehouses. Chapter 4 describes an example of a service based on the presented approach for mapping short reads on the Hadoop cluster.

2 Characteristics and Requirements for DIS

Consider typical requirements for DIS that represent remotely available services for solving a certain class of problems with a large amount of input data. Such services should provide remote interfaces, usually in the form of a web user interface and application programming interface (API). The interface must allow the user to specify the input datasets and parameters of the problem being solved in terms of subject area.

Proceedings of the XIX International Conference "Data Analytics and Management in Data Intensive Domains" (DAMDID/RCDL'2017), Moscow, Russia, October 10–13, 2017

DIS must use high-performance and scalable (normally distributed) implementations of data analysis algorithms, requiring appropriate computing infrastructure for data processing and storage. Such infrastructure is generally represented by one or more computing clusters running Hadoop platform or a similar technology. DIS must translate the user request into one or more computing jobs that are submitted on a cluster and use scalable implementations (e. g., based on MapReduce) of perspective algorithms.

The user must be able to pass arbitrary input data to DIS. If the data is initially located on the user's computer or external storage resource (e. g., a data repository) DIS must implement the transfer of data over a network to the used cluster. When transferring large amounts of data it is important to ensure the maximum transfer rate and automatic failover. Since the process of working with big data is often exploratory, requiring multiple invocations of DIS, the service should support reuse of data loaded to the cluster. In order to optimize the use of network DIS must also cache frequently used datasets on the cluster. Data transfer functions can also be implemented as separate auxiliary services.

Importantly, DIS may operate separately from computing resources used for real data processing. DIS can use multiple resources, that can be situated at different locations. It is also possible that the service uses the resources provided by the user. In such cases it is important for reasons of efficiency to avoid passing the input data from the user to the resource through the service and to transmit the data directly.

In practice, the data analysis is often a multi-step process that requires performing different tasks at different stages of the analysis. In such cases, the results produced by one DIS can be passed as the input to another service. If these services use different resources, there also arises a problem of data transmission between resources. In general DIS should allow the user to download the output to his computer or an external resource, as well as to transfer the data directly to another service. In addition, DIS may provide additional functionality for remote data preview and visualization. These functions may also be implemented as separate auxiliary services.

DIS must support the simultaneous use by multiple users. This requires the protection of user data, resource distribution between users and isolation of computational processes. In the case of cloud infrastructure, DIS must also manage dynamic allocation and deallocation of resources in the cloud, according to the current load.

3 Implementation of DIS with Everest

Everest [7, 8] is a web-based distributed computing platform. It provides users with tools to quickly publish and share computing applications as services. The platform also manages execution of applications on external computing resources attached by users. In contrast to traditional distributed computing platforms, Everest implements the Platform as a Service (PaaS) model by providing its functionality via remote web and programming interfaces. A single instance of the platform can be accessed by many users in order to create, run and share applications with each other. The platform implements integration with servers and computing clusters using an agent that runs on the resource side and plays the role of mediator between the platform and resources. The platform is publicly available online to interested users [8].

The advantage of using Everest platform to create DIS is the availability of ready-made tools for rapid deployment of computational services and integration with computing resources that do not require a separate installation of the platform. At the same time, since the platform was originally created to support services with a small amount of data, the effective implementation of DIS on the base of Everest requires a number of improvements. In particular, it is necessary to implement support of direct data transfers from external storage to the resource and vice versa, bypassing the platform. In addition, it is required to implement the integration of the agent with the components of Hadoop platform platform or similar technology used for data storage and processing on the cluster.

Figure 1 presents the proposed scheme of implementation of DIS on the base of Everest platform and existing Hadoop cluster. Consider the scenario of using the service, which includes the following steps marked in the figure.

In step 1, the user uploads the data of interest to some available on the network or selects data already present in the storage. This storage can be represented by cloud services (Dropbox, Google Drive, etc.), scientific data repositories (Dataverse, FigShare, Zenodo, etc.), specialized databases (for example, 1000 Genomes Project), grid services or file servers (HTTP, FTP, GridFTP, rsync protocols). A wide range of existing storage facilities makes the task of integrating DIS with them more important, in comparison with the duplication of their functionality in the service itself. Note that the user's computer can also act as a data store. In this case, the user needs to deploy a software that provides network access to the user's files. The experience of implementing such software to ensure the reliable transfer of scientific data across the network is already available [9].

In step 2, the user prepares and sends a request to the DIS, including a link to the input data and the values of other input parameters required by the service. The passed link should allow downloading the data from the external storage without the user's participation. In some cases, this requires that the user first supply the service with access credentials to the storage, such as an OAuth token or a proxy certificate.

In step 3, based on the user request, the service generates a computational task and sends it to the agent located on the resource used by the service. Together with the task description, the service sends to an agent a link to the input data. As shown in the figure, when sending a task from the service to the agent, the code of the software implementation used for data processing can also be transferred.

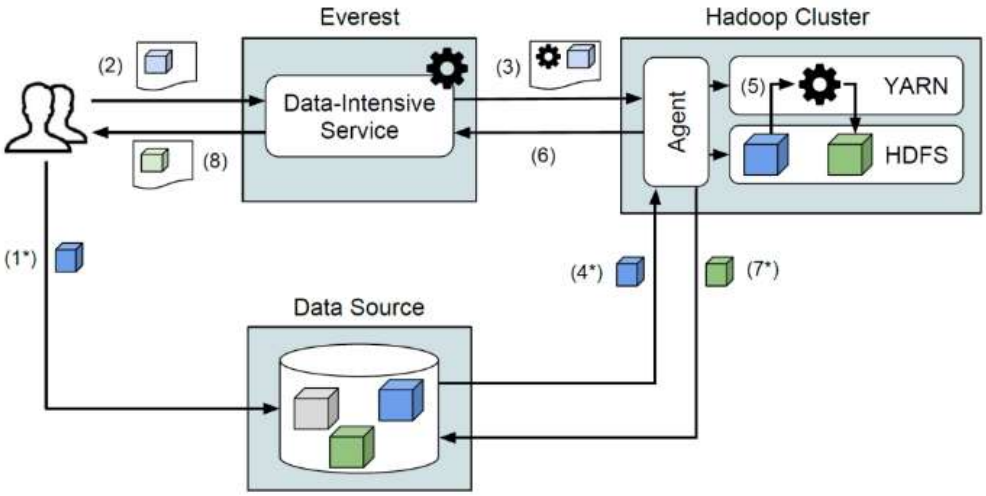


Figure 1 Implementation of DIS on the base of Everest platform and Hadoop cluster

The Hadoop and Spark platforms, most commonly used for distributed data processing, use the Java, Scala, and Python languages for implementation of data processing algorithms. Unlike C and Fortran, often used in scientific parallel applications, programs in these languages can be relatively easily transferred from one cluster to another, including their dependencies, without the need for compilation. This opens the possibility for implementation of services on the basis of already created programs and libraries for Hadoop and Spark, which can be used in conjunction with an arbitrary cluster specified by the user. This model significantly simplifies the publication and reuse of developments in this field, without requiring the owner of the service to provide their own resources. This also avoids the multiple implementations of services that use a single program with different resources.

In step 4, the agent downloads input data from the external storage to the local cluster. To implement this step, it is planned to add support for loading data from major types of repositories and storage. Currently the basic support for downloading files via HTTP and FTP protocols, as well as an experimental integration with Dropbox and Dataverse repository are implemented. The downloaded data is placed in the Hadoop Distributed File System (HDFS) on the cluster, where it can be accessed by the program launched in the next step.

In step 5, the agent runs the program specified in the task description on the given input data. The launch is performed through the cluster resource manager such as Yet Another Resource Negotiator (YARN), a component of the Hadoop platform that supports the launch of MapReduce and Spark programs. A special adapter was implemented in order to support interaction of Everest agent with YARN, similar in function to the previously created adapters for integration with HPC batch schedulers. After the launch, the agent monitors the status of the corresponding job (the application in terms of YARN) and broadcasts the progress information to the service (step 6), which in turn displays this information to the user through the web interface. Upon completion of the program, the agent transmits to the service the output files (of small size) and the final status of the job.

If a large amount of data is produced as a result of the program execution, the agent must support direct network transfer of this data to the user specified external storage (step 7). The information required for this must be transmitted by the user when sending a request to the service in step 2. At the moment, the upload of output data to the specified FTP server or Dropbox folder is implemented.

In step 8, when the request is processed, the service sends the results to the user as a set of output parameters and links to the output files. Some of these files can be stored by the service itself (for example, the program execution log), and some of them can be stored on a cluster or located in an external storage.

Note that the steps 1, 4 and 7, marked with an asterisk and associated with the transmission of data over the network, are not always required or may be omitted. For example, step 1 is not required if the data is already in an external storage or on a cluster, which is true for frequently used data sets. Step 4 can be skipped if the data has already been downloaded to the cluster by the agent or manually by the administrator. To do this, the agent must store information about the downloaded data and cache it for reuse. Step 7 is not required if the received data is an intermediate result and will be submitted as an input to another service using the same cluster. Taking into account these cases can significantly reduce the amount of data transferred across the network and, thus, speedup the processing of requests.

Let us briefly consider security issues. Since the service users can not modify the code of the program launched by the service on the cluster, the risk of unauthorized access to data of other users is minimized.

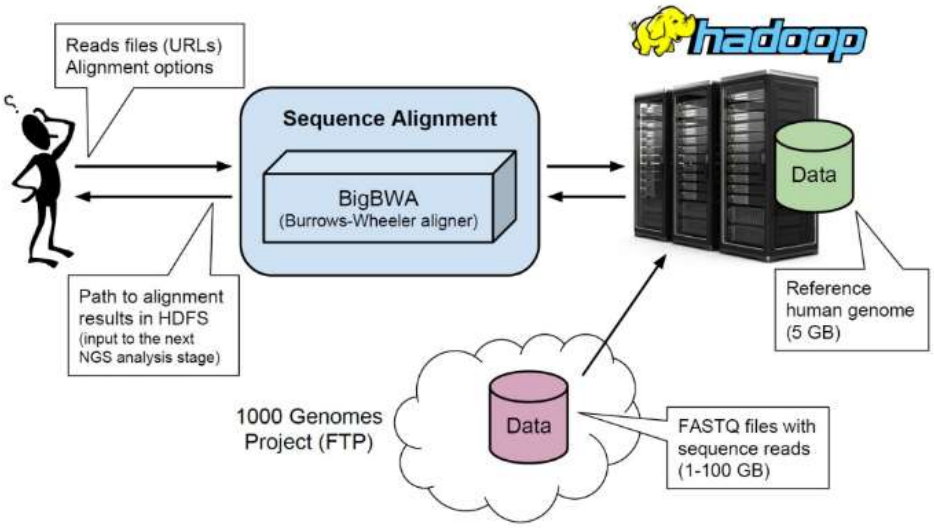


Figure 2 Implementation of DIS for mapping short readings

When implementing data caching on a cluster, the agent must also limit the re-use of confidential data only by the user who originally provided this data. As for the distribution of cluster resources between users and the isolation of computing processes, these functions are already implemented in the YARN manager.

Although the approach described in this section implies the use of the Hadoop platform, it can be easily adapted to any other big data storage and processing platform.

4 Example DIS Implementation

To demonstrate the described approach, a prototype service was implemented on Everest platform for mapping short readings, one of the basic problems of analyzing the results of the next generation sequencing (NGS) in the bioinformatics domain. This task usually represents the initial and the most computationally intensive stage of the NGS data analysis pipeline, characterized by large volumes of input and output data. The basic scheme of the service implementation in presented in Figure 2.

The service requires one or two (paired) files with reads in the FASTQ format to be provided by a user. The size of these files in compressed form is usually several gigabytes. The public repository of the 1000 Genomes Project was chosen as the main input data storage. This repository provides the ability to download data from the dedicated FTP server where both short reads and reference genomes necessary for solving the mapping problem are available. Therefore the files are provided to the service as links to this or any other FTP server.

To solve the mapping problem, the BigBWA tool [10] was used, which implements the parallel execution of the well-known BWA package (Burrows-Wheeler aligner) in the MapReduce paradigm on the Hadoop cluster. When accessing the service, the user can select one of the mapping algorithms implemented in the BWA package. Additional fine-tuning of the algorithm parameters is currently not implemented. Also, all launches use a fixed reference human genome of about 5 GB in size preloaded on the cluster. The total amount of input data of the problem on test runs was about 10–15 GB.

Upon the request submission, in accordance with the scheme described in Chapter 3, the direct downloading of the read files from the FTP server to the Hadoop cluster takes place. After downloading, the files are uncompresses and converted to the format used by BigBWA. The downloaded files are cached and, if the file link in the request matches the already downloaded one, the data loading step is skipped. After the data is loaded, the MapReduce job is launched with the BigBWA tool.

At the end of the job execution, the service returns to the user the path to the file with the mapping results on the cluster. This approach was chosen because, as noted earlier, reads mapping is only the initial step in the analysis of NGS data. Therefore, in practice, these results will usually be immediately passed as an input to another service in the data processing pipeline. At the user's request, the mapping results can also be uploaded to an external FTP server. The output data on the test runs was about 5-10 GB in the SAM format. In the future, it is planned to convert the results into a more compact BAM format.

The solution of the reads mapping problem on the Hadoop cluster via the created service allowed to significantly reduce the data processing time. For example, the launch of the BWA package for mapping on two reads on a single server in 4 threads took more than an hour, while the similar launch through a service (28 map-tasks) took about 10 minutes.

5 Conclusion

In this paper, we considered the characteristic features and requirements for the implementation of data-intensive services for working with large data sets. An approach to the implementation of these services based on the Everest platform, initially focused on the creation of computing services with a small amount of data, is proposed. A distinctive feature of this approach, in comparison with commercial cloud solutions, is support for the rapid implementation of services based on existing computing resources and data repositories. An example of a created service that implements the analysis of next-generation sequencing data on the Hadoop cluster is described.

Besides further development of the individual elements of the described approach, future work will focus on remaining challenges. For instance, many existing data repositories are not well prepared for immediate use and require considerable information integration efforts. There is also an increasing demand for processing of data streams. We plan to investigate the use of data integration and stream processing frameworks within the proposed approach to address these issues. We also plan to evaluate the proposed approach on case study applications using larger data sets or combining data from multiple repositories.

Acknowledgements

This work is supported by the Russian Science Foundation (project No. 16-11-10352).

References

[1] Madduri, R. et al.: Experiences Building Globus Genomics: A Next-generation Sequencing Analysis Service Using Galaxy, Globus, and Amazon Web Services. Concurrency and Computation: Practice and Experience, 26 (13), pp. 2266-2279 (2014)

[2] Madduri, R. et al." PDACS: A Portal for Data Analysis Services for Cosmological Simulations. Computing in Science & Engineering, 17 (5), pp. 18-26 (2015)

[3] Ekanayake, J., Pallickara, S., Fox, G." MapReduce for Data Intensive Scientific Analyses. 2008 IEEE Int. Conf. on eScience (eScience'08). IEEE, pp. 277-284 (2008)

[4] Zhang, Z. et al.: Scientific Computing Meets Big Data Technology: An Astronomy Use Case. 2015 IEEE Int. Conf. on Big Data. IEEE, pp. 918-927 (2015)

[5] Nothaft, F.A. et al.: Rethinking Data-intensive Science Using Scalable Analytics Systems. ACM SIGMOD Int. Conf. on Management of Data. ACM, pp. 631-646 (2015)

[6] Foster, I., Kesselman, C. (ed.). The Grid 2: Blueprint for a New Computing Infrastructure. Elsevier (2003)

[7] Sukhoroslov, O., Volkov, S., Afanasiev, A.A.: Web-based Platform for Publication and Distributed Execution of Computing Applications. 14th Int. Symposium on Parallel and Distributed Computing (ISPDC), pp. 175-184 (2015)

[8] Everest. http://everest.distcomp.org/

[9] Chard, K., Tuecke, S., Foster, I.: Efficient and Secure Transfer, Synchronization, and Sharing of Big Data. Cloud Computing. IEEE, 1 (3), pp. 46-5 (2014).

[10] Abu´ın, J.M. et al.: BigBWA: Approaching the Burrows–Wheeler Aligner to Big Data Technologies. Bioinformatics (2015). doi:10.1093/bioinformatics/btv506

Специализированные инфраструктуры в ОИИД 2

Special-purpose DID infrastructures 2

An Approach to Data Mining Inside PostgreSQL Based on Parallel Implementation of UDFs

© Timofey Rechkalov © Mikhail Zymbler

South Ural State University,
Chelyabinsk, Russia

trechkalov@yandex.ru mzym@susu.ru

Abstract. Relational DBMSs remain the most popular tool for data processing. However, most of stand-alone data mining packages process flat files outside a DBMS. In-database data mining avoids export-import data/results bottleneck as opposed to use stand-alone mining packages and keeps all the benefits provided by DBMS. The paper describes an approach to data mining inside PostgreSQL based on parallel implementation of user-defined functions (UDFs) for modern Intel many-core platforms. The UDF performs a single mining task on data from the specified table and produces a resulting table. The UDF is organized as a wrapper of an appropriate mining algorithm, which is implemented in C language and is parallelized based on OpenMP technology and thread-level parallelism. The library of such UDFs supports a cache of precomputed mining structures to reduce costs of computations. We compare performance of our approach with *R* data mining package, and experiments show efficiency of the proposed approach.

Keywords: data mining, in-database analytics, PostgreSQL, thread-level parallelism, OpenMP.

1 Introduction

Currently relational DBMSs remain the most popular facility for storing, updating and querying structured data. At the same time, most of data mining algorithms suppose processing of flat file(s) outside a DBMS. However, exporting data sets and importing of mining results impede analysis of large databases outside a DBMS [18]. In addition to avoiding export-import bottleneck, an approach to data mining inside a DBMS provides many benefits for the end-user like query optimization, data consistency and security, etc.

Existing approaches to integrating data mining with relational DBMSs include special data mining languages and SQL extensions, implementation of mining algorithms in plain SQL and user-defined functions (UDFs) implemented in high-level language like C++. The latter approach could serve as a subject of applying parallel processing on modern many-core platforms.

In this paper, we present an approach to data mining inside PostgreSQL open-source DBMS exploiting capabilities of modern Intel MIC (Many Integrated Core) [2] platform. Our approach supposes a library of UDFs where each one of them performs a single mining task on data from the specified table and produces a resulting table. The UDF is organized as a wrapper of an appropriate mining algorithm, which is implemented in C language and is parallelized for Intel MIC platform by OpenMP technology and thread-level parallelism.

The paper is structured as follows. We describe the proposed approach in the Section 2. The results of experimental evaluation of our approach are given in Section 3. Section 4 briefly discusses related works. Section 5 contains summarizing comments and directions for future research.

Proceedings of the XIX International Conference "Data Analytics and Management in Data Intensive Domains" (DAMDID/RCDL'2017), Moscow, Russia, October 10–13, 2017

2 Embedding of data mining functions into PostgreSQL

2.1 Motivation example

Our approach is aimed to provide a database application programmer with the library of data mining functions, which could be run inside DBMS as it shown in Fig. 1.

```
#include <libpq-fe.h> // API of PostgreSQL
#include "pgmining.h" // API of pgMining library

void main(void)
{
  char * inpTable = "points";
  char * outTable = "clusters";
  int dimension = 3;
  int k = 5;
  float Eps = 0.1;
  char * conninfo = "user=postgres port=5432 host=localhost";

  PGconn * conn = PQconnectdb(conninfo);
  pgPAM(conn, inpTable, dimension, k, Eps, outTable);
  PQexec(conn, "SELECT * FROM clusters;");
  PQfinish(conn);
}
```

Figure 1 An example of using data mining function inside PostgreSQL

In this example the mining function performs clustering by Partitioning Around Medoids (PAM) [8] algorithm for the data points from the specified input table and saves results in output table (with respect to the specified number of the input table's columns, number of clusters and accuracy). An application programmer is not obliged to export data to be mined from DBMS and import mining results back. At the same time here PAM encapsulates parallel implementation [24] based on OpenMP technology and thread-level parallelism.

2.2 Component structure

Fig. 2 depicts the component structure of our approach. The *pgMining* is a library of data mining functions each one of them is to be run inside PostgreSQL. The *mcMining* is a library that exports data mining functions, which are parallelized for modern many-core platforms and are subject of wrapping by the respective functions from *pgMining* library. Implementation of *pgMining* library uses PostgreSQL's SPI (Server Programming Interface), which provides low level functions for data access.

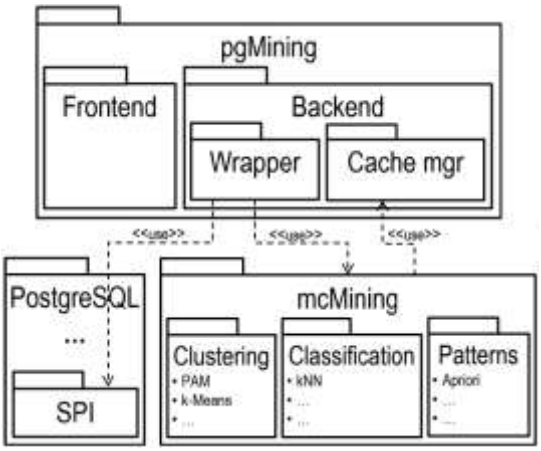


Figure 2 Component structure of the proposed approach

The *pgMining* library consists of two following subsystems, namely *Frontend* and *Backend*, where the former provides presentation layer and the latter – data access layer of concerns for an application programmer.

The *Frontend* provides a set of functions for mining inside PostgreSQL. Each function performs a single mining task (e.g. clustering, classification, search patterns, etc.) and produces a resulting table.

The *Backend* consists of two modules, namely *Wrapper* and *Cache manager*. The *Wrapper* provides functions that serve as envelopes for the respective mining functions from *mcMining* library. The *Cache manager* supports cache of precomputed mining structures to reduce costs of computations.

The *mcMining* library provides a set of functions to solve various data mining tasks in main memory and exploits capabilities of Intel many-core platforms.

2.3 Frontend

An example of *Frontend*'s function is given in Fig. 3. Such a function connects to PostgreSQL, carries out some mining task and returns exit code (0 in case of success, otherwise negative error code). As a side effect, the function creates a table with mining results. The function's mandatory parameters are ID of PostgreSQL connection, name of the input table, name of the output table and number of first left columns in input table containing data to be mined. The rest parameters are specific to the task (e.g. number of clusters, accuracy, etc.).

```
// PAM clustering inside PostgreSQL
// Returns 0 in case of success or negative error code.
int pgPAM(
  PGconn * conn,   // ID of PostgreSQL connection
  char * inpTable, // Name of input table
  int dimension,   // Number of coordinates in data point
  int k,           // Number of clusters
  float Eps,       // Accuracy of computations
  char * outTable)// Name of output table
{
  PQexec(conn, "CREATE OR REPLACE FUNCTION
    wrap_pgPAM(text, integer, integer, real) RETURNS text AS
      'pgmining', 'wrap_pgPAM' LANGUAGE C STRICT;");
  PQexec(conn, "CREATE %s TABLE IF NOT EXISTS (data text)",
    outTable);
  return PQexec(conn, "INSERT INTO %s
    SELECT wrap_pgPAM(%s, %d, %d, %f);",
      outTable, inpTable, dimension, k, Eps);
}
```

Figure 3 Interface and implementation schema of function from *Frontend*

In fact, *Frontend*'s function wraps the respective UDF from *Backend*, which is loaded into PostgreSQL and executed as "`INSERT INTO … SELECT …`" query to save mining results in the specified table.

2.4 Backend

Fig. 4 depicts an example of *Wrapper*'s function. Such a function is an UDF, which wraps a parallelized mining function from *mcMining* and performs as follows. Firstly, the function parses its input to form parameters to call *mcMining* function with. After that, the function checks if input table and/or auxiliary mining structures are in the cache maintained by *Cache manager* and then load them if not. Finally, call of *mcMining* function with appropriate parameters is performed.

```
#include "postgres.h"

// Wrapper for PAM clustering inside PostgreSQL
// Returns 0 in case of success or negative error code.
Datum wrap_pgPAM(PG_FUNCTION_ARGS)
{
  // Extract parameters of the algorithm
  char * inpTable = text_to_cstring(PG_GETARG_TEXT_P(0));
  int dimension = PG_GETARG_INT32(1);
  int k = PG_GETARG_INT32(2);
  float Eps = PG_GETARG_FLOAT4(3);
  int N;
  // Check if mining structure is in the cache
  void * distMatrix = cache_getObject(
    strcat(inpTable, "_distMatrix"));
  if (distMatrix == NULL) {
    // Check if input table is in the cache
    void * inpData = cache_getObject(inpTable);
    if (inpData == NULL) {
      // Allocate memory and load input table to cache
      inpData = (float4 *) palloc(dimension*sizeof(float4));
      wrap_tabRead(inpData, inpTable, dimension, &N);
      cache_putObject(inpTable, inpData, sizeof(inpData));
    }
    distMatrix = mcCalcMatrix(inpData, dimension, N);
    cache_putObject(strcat(inpTable, "_distMatrix"),
      distMatrix, sizeof(distMatrix));
  }
  // Perform clustering
  mcPAM_res * outData = mcPAM_resCreate();
  mcPAM(N, k, Eps, outData, distMatrix);
  // Write results to the output table
  PG_RETURN_TEXT(data2String(outData));
}
```

Figure 4 Interface and implementation schema of function from *Backend*

The *Cache manager* provides buffer pool to store precomputed mining structures. Distance matrix is a typical example of mining structure to be saved in cache. Indeed, distance matrix $A=(a_{ij})$ stores distances between

each pair of a_i and a_j elements in input data set. Being precomputed once, distance matrix could be used many times to perform clustering or kNN-based classification with various parameters (e.g. number of clusters, number of neighbors, accuracy, etc.).

```
// Load an object to cache.
// Returns 0 in case of success or negative error code.
int cache_putObject(
  char * objID, // ID of the object
  void * data, // Pointer to data buffer
  int size);   // Size of data

// Search an object with given ID in cache.
// Returns pointer to object in case of success or NULL.
void * cache_getObject(char * objID);
```

Figure 5 Interface of *Cache manager* module

The *Cache manager* exports the following two basic functions depicted in Fig. 5. The `putObject` function loads a mining structure specified by its ID, buffer pointer and size into cache. The `getObject` searches in cache for an object with the given ID. An ID of mining structure is a string, which is made as concatenation of input table's name and object's informational string (e.g. "`_distMatrix`").

2.5 Library of parallel many-core algorithms

Fig. 6 gives an example of function from *mcMining* library. Such a function encapsulates parallel implementation through OpenMP technology and thread-level parallelism for Intel many-core platforms.

```
// PAM clustering parallelized for Intel many-core platform.
// Returns 0 in case of success or negative error code.
int mcPAM(
    int N,          // Number of data points
    int k,          // Number of clusters
    float Eps,      // Accuracy of computations
    void * outData, // Array of output centroids
    void * distMatrix); // Precomputed distance matrix
```

Figure 6 Interface of function from *mcMining* library

In this example, we use Partition Around Medoids (PAM) [8] clustering algorithm, which is used in a wide spectrum of applications where minimal sensitivity to noise data is required. The PAM provides such a property since it represents cluster centers by points of input data set (*medoids*).

The PAM firstly calculates distance matrix for the given data points. Then in the BUILD phase, an initial clustering is obtained by the successive selection of medoids until the required number of clusters have been found. Next, in the SWAP phase the algorithm attempts to improve clustering in accordance with an objective function. However, for large and high-dimensional datasets PAM's computations are very costly.

In our previous research [24], we parallelize PAM for Intel Xeon CPU and Intel Xeon Phi coprocessor. In order to perform best on Intel many-core platforms the PAM's parallel version exploits modifications of loops to provide vectorization of calculations and chunk-by-chunk data processing to decrease number of cache misses.

3 Experimental evaluation

3.1 Hardware, datasets and goals of experiments

To evaluate the developed approach, we performed experiments on the Tornado SUSU supercomputer [9] whose node provides two different platforms, namely Intel Xeon CPU and Intel Xeon Phi coprocessor (cf. Tab. 1 for the specifications).

Table 1 Specifications of hardware

Specifications	**CPU**	**Coprocessor**
Model, Intel Xeon	X5680	Phi SE10X
Cores	2×6	61
Frequency, GHz	3.33	1.1
Threads per core	2	4
Peak performance, TFLOPS	0.371	1.076
Memory, Gb	24	18
Cache, Mb	12	30.5

In the experiments, we used datasets with the characteristics depicted in Tab. 2.

Table 2 Summary of datasets used in experiments

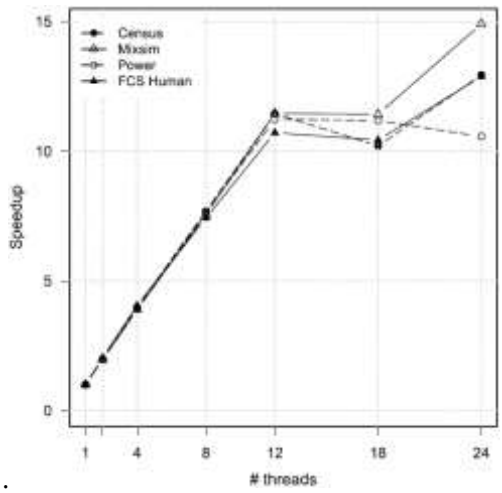
Dataset	**dimen-sion**	**# clusters**	**# data points**, $\times 2^{10}$
FCS Human [3]	423	10	18
MixSim [13]	5	10	35
US Census [12]	67	10	35
Power Consumption [10]	3	10	35

In the experiments, we studied the following aspects of the developed approach. Firstly, we investigated the speedup of *mcPAM* function to understand its scalability on both platforms depending on number of threads employed. Secondly, we evaluated the runtime of *mcPAM* function to understand how the performance on both platforms depends on number of data points and what benefits could we derive from precomputations of the distance matrix. Finally, we compared the performance of *pgPAM* function with implementation of PAM algorithm from *R* data mining package [13].

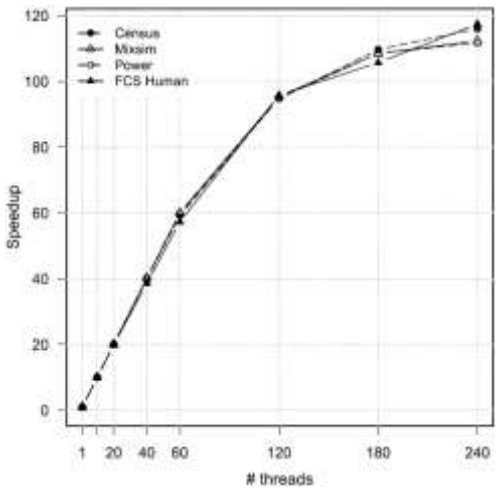
3.2 Results of experiments

The results of the first series of experiments on *mcPAM* speedup are depicted in Fig. 7. On both platforms, *mcPAM*'s speedup is close to linear, when the number of threads matches the number of physical cores the algorithm is running on (i.e. 12 cores for Intel Xeon and 60 cores for Intel Xeon Phi, respectively).

Speedup becomes sub-linear when the algorithm uses more than one thread per physical core. The *mcPAM* achieves up to 15× and 120× speedup on Intel Xeon and Intel Xeon Phi, respectively. Summing up, *mcPAM* demonstrates good scalability on both platforms

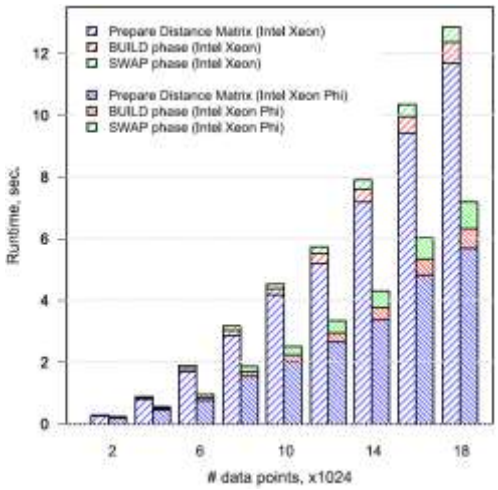


(a) Intel Xeon CPU

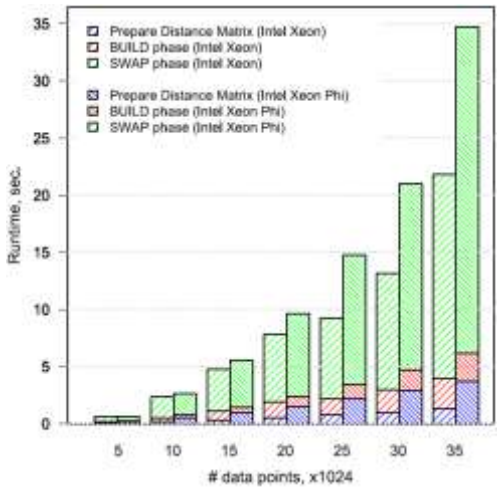


(b) Intel Xeon Phi coprocessor

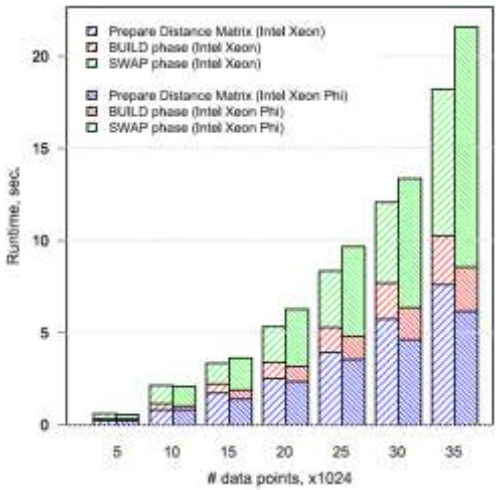
Figure 7 Speedup of the *mcPAM* function



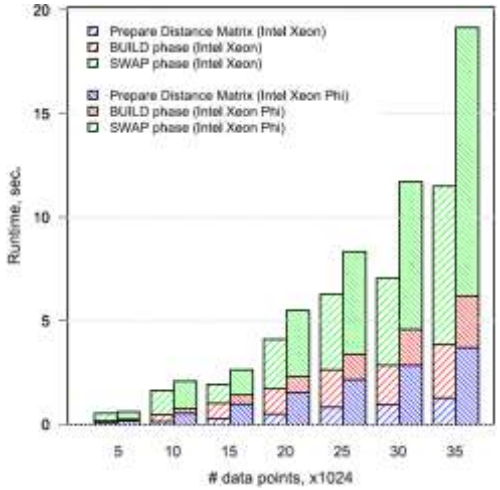
(a) FCS Human dataset



(b) MixSim dataset

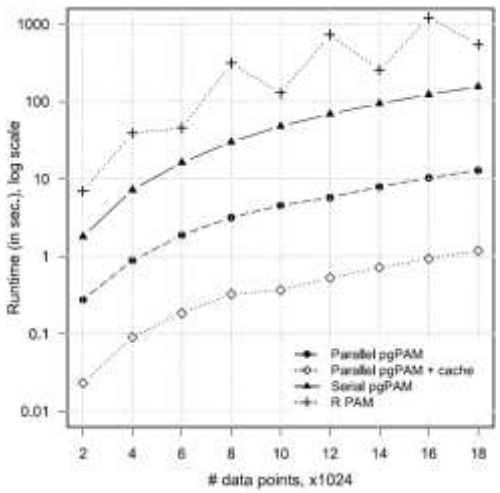


(c) US Census dataset

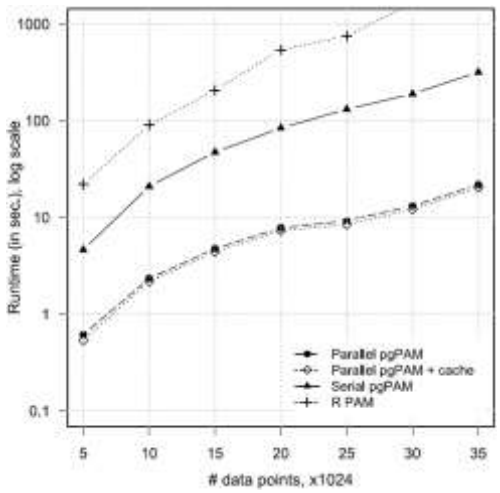


(d) Power Consumption dataset

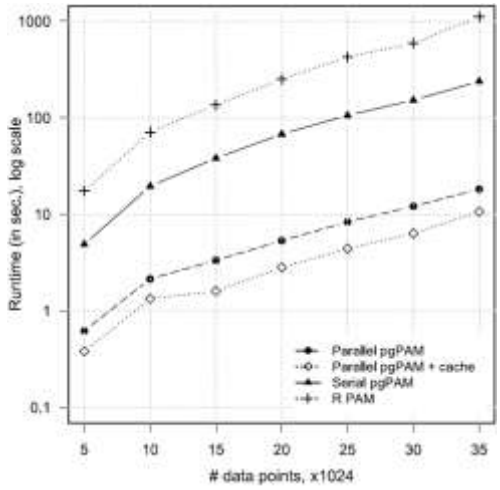
Figure 8 Performance of the *mcPAM* function



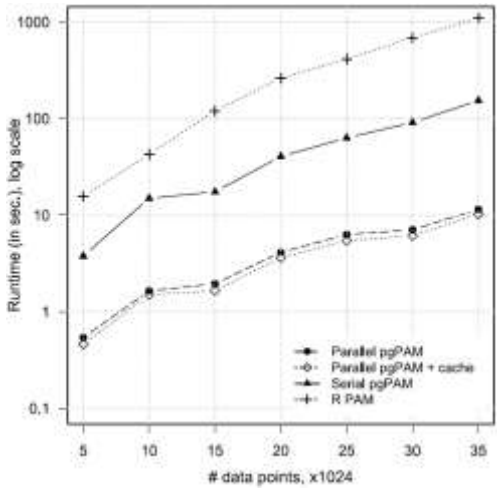
(a) FCS Human dataset



(b) MixSim dataset



(c) US Census dataset



(d) Power Consumption dataset

Figure 9 Performance of the *pgPAM* function (on Intel Xeon)

Fig. 8 shows the results of the second series of experiments on *mcPAM* performance. As was seen, PAM's SWAP phase is performed better on Intel Xeon while BUILD phase performance is equal for both platforms.

Overall performance is better on Intel Xeon Phi than Intel Xeon when the algorithm deals with big dimensionality dataset due to possibility of intensive vectorization in calculations of distance matrix. Since calculations of distance matrix take from 15 to 80 percent of overall runtime, we can derive substantial benefits from caching of the distance matrix.

The results of the third series of experiments on comparison performance of *pgPAM* and PAM from *R* data mining package are illustrated in Fig. 9. We carried out these series of experiments on Intel Xeon platform only due to the following reason. Running PostgreSQL on Intel MIC platform demands Intel Xeon Phi Knights Landing (KNL), which is the next generation product from Intel and is bootable device. However, Intel Xeon Phi KNL is not available yet at Tornado SUSU supercomputer. We plan to perform this study as further research.

We can see that *pgPAM* significantly overtakes *R*'s PAM in both cases when one thread or the maximum number of threads are employed. Caching of distance matrix improves the performance up to 80 percent of overall runtime (in case of high-dimensional dataset).

4 Related work

The problem of integrating data analytics with relational DBMSs has been studied since data mining research originates.

Data mining query languages include DMQL [5], MSQL [7], MINE RULE operator [14] and Microsoft's DMX [28].

There are many *SQL implementations* of data mining algorithms. SQL versions of classical clustering algorithms include K-Means [16], EM [19], Fuzzy C-Means [15]. SQL versions of association rule mining algorithms include K-Way-Join, Three-Way-Join,

Subquery and Two-Group-Bys [25], Set-oriented Apriori [29], Quiver [23], Propad [27]. Classification includes SQL implementations of decision trees [26], kNN [31] and Bayesian classification [21]. SQL is also successfully used in mining applications for data with "non-relational" nature as graphs, for instance in search for frequent graphs [4], detection of cycles in graph [1], graph partitioning [11, 22], etc.

User-defined functions-based approach. Integration of correlation, linear regression, PCA and clustering into the Teradata DBMS based on UDFs is proposed in [17]. There are two sets of UDFs that work in a single table scan, that is an aggregate UDF to compute summary matrices and a set of scalar UDFs to score data sets. Experiments showed that UDFs are faster than SQL queries and UDFs are more efficient than C++, due to long export times. In [20] UDFs implementing common vector operations were presented and it was shown that UDFs are as efficient as automatically generated SQL queries with arithmetic expressions and queries calling scalar UDFs are significantly more efficient than equivalent queries using SQL aggregations.

In-database mining frameworks. The ATLAS [30] is a framework for in-database analytics, which provides SQL-like database language with user-defined aggregates (UDAs) and table functions. The system's language processor translates ATLAS programs into C++ code, which is then compiled and linked with the database storage manager and user-defined external functions. Authors presented ATLAS-based implementations of several data mining algorithms.

The MADlib [6] is an open source library of in-database analytical algorithms for PostgreSQL. The MADlib is implemented by a big team and provides many methods for supervised learning, unsupervised learning and descriptive statistics. The MADlib exploits UDAs, UDFs, and a sparse matrix C library to provide efficient representations on disk and in memory. As many statistical methods are iterative (i.e. they make many passes over a data set), authors wrote a driver UDF in Python to control iteration in such a way that all large data movement is done within the database engine and its buffer pool.

Comparison. In this paper, we suggest an approach to embedding data mining functions into PostgreSQL. As some methods mentioned above our approach exploits UDFs. The difference from the previous works includes the following. Our approach supposes parallelization of UDFs for many-core platform that current DBMS is running on. All the parallelization details are encapsulated in implementation of the UDF and are hided from the DBMS, so our approach could be ported to some other open-source DBMS (with possible non-trivial but mechanical software development effort). In addition, our approach supposes a special module, which provides a cache of precomputed mining structures and lets UDF know to reuse these structures to reduce costs of computations.

5 Conclusion

In this paper, we touch upon the problem of organizing data mining inside a DBMS. We present an approach to implementation of in-database analytical functions for PostgreSQL that exploits capabilities of modern Intel many-core platforms.

Our approach supposes implementation of two libraries, namely *pgMining* and *mcMining*. The *pgMining*is a library of data mining functions each one of them is to be run inside PostgreSQL. The *mcMining* is a library that exports functions to solve various data mining tasks, which are parallelized for Intel MIC platforms.

The *pgMining* consists of *Frontend* and *Backend* subsystems. The *Frontend*'s function loads an UDF from the *Backend* into PostgreSQL and executes it as "`INSERT INTO … SELECT …`" query to save mining results in a table. The *Backend* consists of *Wrapper* and *Cache manager* modules. The *Wrapper* provides functions that serve as envelopes for the respective *mcMining* mining functions. The *Cache manager* supports cache of precomputed mining structures to reduce costs of computations.

Since our approach assumes hiding details of parallel implementation from PostgreSQL, such an approach could be ported to some other open-source DBMS (with possible non-trivial but mechanical software development effort).

We have evaluated our approach on previously implemented parallel clustering algorithm of *mcMining* library and four real datasets. Experiments showed good speedup and performance of the algorithm as well as our approach derive benefits from caching of precomputed mining structures and overtakes *R* data mining package.

As future work, we plan to implement other mining algorithms for*mcMining* library and conduct experiments on Intel Xeon Phi Knights Landing platform.

Acknowledgement

This work was financially supported by the Russian Foundation for BasicResearch (grant No. 17-07-00463), by Act 211 Government of the Russian Federation (contract No. 02.A03.21.0011) and by the Ministry of education and science of Russian Federation (government order 2.7905.2017/8.9).

References

[1] Balachandran, R., Padmanabhan, S., Chakra-varthy, S.: Enhanced DB-Subdue: Supporting Subtle Aspects of Graph Mining Using a Relational Approach. In: W.K. Ng, M. Kitsuregawa, J. Li, K. Chang (eds.) Advances in Knowledge Discovery and Data Mining, 10th Pacific-Asia Conf., PAKDD 2006, Singapore, April 9–12, 2006, Proc., Lecture Notes in Computer Science, 3918, pp. 673-678. Springer (2006). doi:10.1007/ 11731139 77

[2] Duran, A., Klemm, M.: The Intel Many Integrated Core Architecture. In: W.W. Smari, V. Zeljkovic (eds.) HPCS, pp. 365-366. IEEE (2012)

[3] Engreitz, J.M., Jr., B.J.D., Marshall, J.J., Altman, R.B.: Independent Component Analysis: Mining Microarray Data for Fundamental Human Gene Expression Modules. J. of Biomedical Informatics. 43 (6), pp. 932-944 (2010)

[4] Garcia, W., Ordonez, C., Zhao, K., Chen, P.: Efficient Algorithms Based on Relational Queries to Mine Frequent Graphs. In: A. Nica, A.S. Varde (eds.) Proc. of the Third Ph.D. Workshop on Information and Knowledge Management, PIKM 2010, Toronto, Ontario, Canada, October 30, 2010, pp. 17-24. ACM (2010). doi:10.1145/ 1871902.1871906

[5] Han, J., Fu, Y., Wang, W., Chiang, J., Gong, W., Koperski, K., Li, D., Lu, Y., Rajan, A., Stefanovic, N., Xia, B., Zaiane, O.R.: Dbminer: A System for Mining Knowledge in Large Relational Databases. In: E. Simoudis, J. Han, U.M. Fayyad (eds.) Proc. of the Second Int. Conf. on Knowledge Discovery and Data Mining (KDD-96), Portland, Oregon, USA, pp. 250-255. AAAI Press (1996)

[6] Hellerstein, J.M., Re, C., Schoppmann, F., Wang, D.Z., Fratkin, E., Gorajek, A., Ng, K.S., Welton, C., Feng, X., Li, K., Kumar, A.: The MADlib Analytics Library or MAD Skills, the SQL. PVLDB 5(12), pp. 1700-1711 (2012)

[7] Imielinski, T., Virmani, A.: MSQL: A Query Language for Database Mining. Data Min. Knowl. Discov. 3 (4), pp. 373-408 (1999). doi: 10.1023/A:1009816913055

[8] Kaufman, L., Rousseeuw, P.J.: Finding Groups in Data: An Introduction to Cluster Analysis. John Wiley (1990)

[9] Kostenetskiy, P., Safonov, A.: SUSU Supercomputer Resources. In: L. Sokolinsky, I. Starodubov (eds.) PCT'2016, Int. Scientific Conf. on Parallel Computational Technologies, Arkhangelsk, Russia, March 29–31, 2016, pp. 561-573. CEUR Workshop Proceedings. 1576 (2016)

[10] Lichman, M.: UCI Machine Learning Repository [http://archive.ics.uci.edu/ml/datasets/individual+ household+electric+power+consumption]. Irvine, CA: University of California, School of Information and Computer Science (2013)

[11] McCaffrey, J.D.: A Hybrid System for Analyzing Very Large Graphs. In: S. Latifi (ed.) Ninth Int. Conf. on Information Technology: New Generations, ITNG 2012, Las Vegas, Nevada, USA, 16–18 April, 2012, pp. 253-257. IEEE Computer Society (2012). doi:10.1109/ITNG.2012.43

[12] Meek, C., Thiesson, B., Heckerman, D.: The Learning-curve Sampling Method Applied to Model-based Clustering. J. of Machine Learning Research. 2, pp. 397-418 (2002)

[13] Melnykov, V., Chen, W.C., Maitra, R.: Mixsim: An R Package for Simulating Data to Study Performance of Clustering Algorithms. J. of Statistical Software, Articles 51 (12), pp. 1-25 (2012). doi:10.18637/jss.v051.i12

[14] Meo, R., Psaila, G., Ceri, S.: A New SQL-like Operator for Mining Association Rules. In: T.M. Vijayaraman, A.P. Buchmann, C. Mohan, N.L. Sarda (eds.) VLDB'96, Proc. of 22th Int. Conf. on Very Large Data Bases, September 3–6, 1996, Mumbai (Bombay), India, pp. 122-133. Morgan Kaufmann (1996)

[15] Miniakhmetov, R., Zymbler, M.: Integration of Fuzzy c-means Clustering Algorithm with PostgreSQL Database Management System. Numerical Methods and Programming 13 (2(26)), pp. 46-52 (2012) (in Russian)

[16] Ordonez, C.: Integrating k-means Clustering with a Relational DBMS Using SQL. IEEE Trans. Knowl. Data Eng. 18 (2), pp. 188-201 (2006). doi:10.1109/TKDE.2006.31

[17] Ordonez, C.: Building Statistical Models and Scoring with UDFs. In: C.Y. Chan, B.C. Ooi, A. Zhou (eds.) Proc. of the ACM SIGMOD Int. Conf. on Management of Data, Beijing, China, June 12–14, 2007, pp. 1005-1016. ACM (2007). doi:10.1145/1247480.1247599

[18] Ordonez, C.: Statistical Model Computation with UDFs. IEEE Trans. Knowl. Data Eng. 22 (12), pp. 1752-1765 (2010). doi:10.1109/TKDE.2010.44

[19] Ordonez, C., Cereghini, P.: SQLEM: Fast Clustering in SQL Using the EM Algorithm. In: W. Chen, J.F. Naughton, P.A. Bernstein (eds.) Proc. of the 2000 ACM SIGMOD Int. Conf. on Management of Data, May 16–18, 2000, Dallas, Texas, USA, pp. 559-570. ACM (2000). doi: 10.1145/ 342009.335468

[20] Ordonez, C., Garcia-Garcia, J.: Vector and Matrix Operations Programmed with UDFs in a Relational DBMS. In: P.S. Yu, V.J. Tsotras, E.A. Fox, B. Liu (eds.). Proc. of the 2006 ACM CIKM Int. Conf. on Information and Knowledge Management, Arlington, Virginia, USA, November 6–11, 2006, pp. 503-512. ACM (2006). doi:10.1145/ 1183614.1183687

[21] Ordonez, C., Pitchaimalai, S.K.: Bayesian Classifiers Programmed in SQL. IEEE Trans. Knowl. Data Eng. 22 (1), pp. 139-144 (2010). doi: 10.1109/TKDE.2009.127

[22] Pan, C., Zymbler, M.: Very Large Graph Partitioning by Means of Parallel DBMS. In: B. Catania, G. Guerrini, J. Pokorny (eds.) Advances in Databases and Information Systems – 17th East European Conf., ADBIS 2013, Genoa, Italy, September 1–4, 2013. Proc., Lecture Notes in Computer Science, 8133, pp. 388-399. Springer (2013). doi: 10.1007/978-3-642-40683-6 29

[23] Rantzau, R.: Frequent Itemset Discovery with SQL Using Universal Quantification. In: R. Meo, P.L. Lanzi, M. Klemettinen (eds.) Database Support for Data Mining Applications: Discovering Knowledge with Inductive Queries, Lecture Notes in Computer

Science, 2682, pp. 194-213. Springer (2004). doi: 10.1007/ 978-3-540-44497-8 10

[24] Rechkalov, T., Zymbler, M.: Accelerating Medoids-based Clustering with the Intel Many Integrated Core Architecture. In: 9th Int. Conf. on Application of Information and Communication Technologies, AICT 2015, October 14–16, 2015, Rostov-on-Don, Russia. Proceedings, pp. 413-417 (IEEE, 2015). doi:10.1109/ICAICT.2015. 7338591

[25] Sarawagi, S., Thomas, S., Agrawal, R.: Integrating Association Rule Mining with Relational Database systems: Alternatives and Implications. Data Min. Knowl. Discov. 4 (2/3), pp. 89-125 (2000). doi:10.1023/A:1009887712954

[26] Sattler, K., Dunemann, O.: SQL Database Primitives for Decision Tree Classifiers. In: Proc. of the 2001 ACM CIKM Int. Conf. on Information and Knowledge Management, Atlanta, Georgia, USA, November 5–10, 2001, pp. 379-386. ACM (2001). doi:10.1145/502585.502650

[27] Shang, X., Sattler, K., Geist, I.: SQL Based Frequent Pattern Mining with FPGrowth. In: D. Seipel, M. Hanus, U. Geske, O. Bartenstein (eds.) Applications of Declarative Programming and Knowledge Management, 15th Int. Conf. on Applications of Declarative Programming and Knowledge Management, INAP 2004, and 18th Workshop on Logic Programming, WLP 2004, Potsdam, Germany, March 4–6, 2004, Revised Selected Papers, Lecture Notes in Computer Science, 3392, pp. 32-46. Springer (2004). doi: 10.1007/11415763 3

[28] Tang, Z., Maclennan, J., Kim, P.P.: Building Data Mining Solutions with OLE DB for DM and XML for Analysis. SIGMOD Record, 34 (2), pp. 80-85 (2005). doi:10.1145/1083784.1083805

[29] Thomas, S., Chakravarthy, S.: Performance Evaluation and Optimization of Join Queries for Association Rule Mining. In: M.K. Mohania, A.M. Tjoa (eds.) Data Warehousing and Knowledge Discovery, First Int. Conf., DaWaK '99, Florence, Italy, August 30 – September 1, 1999, Proc., Lecture Notes in Computer Science, 1676, pp. 241-250. Springer (1999). doi:10.1007/ 3-540-48298-9 26

[30] Wang, H., Zaniolo, C., Luo, C.: ATLAS: A Small but Complete SQL Extension for Data Mining and Data Streams. In: VLDB, pp. 1113-1116 (2003)

[31] Yao, B., Li, F., Kumar, P.: K Nearest Neighbor Queries and kNN-joins in Large Relational Databases (almost) for Free. In: F. Li, M.M. Moro, S. Ghandeharizadeh, J.R. Haritsa, G. Weikum, M.J. Carey, F. Casati, E.Y. Chang, I. Manolescu, S. Mehrotra, U. Dayal, V.J. Tsotras (eds.) Proc. of the 26th Int. Conf. on Data Engineering, ICDE 2010, March 1–6, 2010, Long Beach, California, USA, pp. 4-15. IEEE Computer Society (2010). doi:10.1109/ICDE.2010.5447837

Вопросы обеспечения информационной безопасности информационных систем, реализующих интенсивное использование данных

© В.Г. Беленков © С.В. Борохов © В.И. Будзко © П.А. Кейер © В.И. Королев

Федеральный исследовательский центр «Информатика и управление» Российской академии наук,

Москва, Россия

vbelenkov@ipiran.ru sborokhov@ipiran.ru vbudzko@ipiran.ru pkeyer@ipiran.ru vkorolev@ipiran.ru

Аннотация. Рассмотрены актуальные вопросы обеспечения информационной безопасности информационных систем, реализующих интенсивное использование данных. Описаны текущее состояние и перспективные направления дальнейших исследований.

Ключевые слова: интенсивное использование данных, большие данные, информационная безопасность.

The Issues of Information Security Provision of Information Systems Used in Data Intensive Domains

© V.G. Belenkov © S.V. Borokhov © V.I. Budzko © P.A. Keyer © V.I. Korolev

Federal Research Center Computer Science and Control of the Russian academy of Sciences, Moscow, Russia

vbelenkov@ipiran.ru sborokhov@ipiran.ru vbudzko@ipiran.ru pkeyer@ipiran.ru vkorolev@ipiran.ru

Abstract. The article discusses topical issues of information security provision of information systems used in Data Intensive Domains. The current state and prospective directions of further research are described.

Keywords: Data Intensive Domains, Big Data, information security**.**

1 Введение

В настоящее время информационные системы, применяемые в различных областях с интенсивным использованием данных (Data Intensive Domains – DID), получают широкое распространение для решения большого круга практических задач. К пониманию целесообразности использования DID-систем в своей деятельности пришли не только крупные коммерческие структуры, но и государственные организации и ведомства, среди которых Федеральная налоговая служба, Банк России и другие [1–4].

Возможности информационных систем, основанных на интенсивном использовании данных, представляются в настоящее время настолько значительными, что некоторые исследователи говорят о возникновении новой, «четвертой парадигмы» науки [5] и «цифровой революции», благодаря которой качество принимаемых машинами решений начинает превосходить качество решений, принимаемых людьми [6]. В 2015 году Gartner Group исключила технологии больших данных из Hype cycle новых технологий (Hype Cycle for Emerging Technologies). Аналитик Gartner Group Betsy Burton объяснила этот факт тем, что «... большие данные распространились в нашей жизни во многих областях и стали частью множества других hype cycles» [7]. В настоящее время DID-системы выявляют мошенников [3], в 10 раз снижают вероятность ошибки первичного диагноза, поставленного врачом [6], дают точные прогнозы результатов выборов [8], позволяют количественно оценивать произошедшие в обществе исторические изменения [8], персонализировать сервис авиаперевозчикам [9] и даже сдают выпускные экзамены в медицинском университете, приобретая тем самым юридическое право лечить людей [6]. Всё это свидетельствует о глубокой интеграции DID-систем в нашу жизнь.

Труды XIX Международной конференции «Аналитика и управление данными в областях с интенсивным использованием данных» (DAMDID/ RCDL'2017), Москва, Россия, 10–13 октября 2017 года

Очевидно, что чем больше люди в своей деятельности будут полагаться на DID-системы, тем более давать полномочия этим системам принимать решения, тем критичнее для общества становятся результаты как непреднамеренных ошибок функционирования, так и реализации злоумышленниками угроз в отношении DID-систем. Кроме этого, отдельной проблемой являются вопросы правового регулирования использования организациями DID-систем и результатов их деятельности. В качества примеров приведём следующие случаи раскрытия конфиденциальной информации, ставшие классическими примерами для DID-систем.

В 2012 году торговая компания Target смогла определить беременность девочки-подростка и разослала соответствующие рекламные буклеты её отцу, таким образом фактически проинформировав его до того, как девочка сама сообщила ему об этом [10–12]. Информация о беременности девочки была получена торговой компанией косвенным путём в результате обработки DID-системой по определению беременности данных о совершаемых ею покупках. В этой ситуации мы имеем дело с тем, что на основании обработки DID-системой данных, являющихся с точки зрения клиентки общедоступными и прямым образом не указывающих на её беременность, была получена информация, относящаяся с точки зрения клиентки к конфиденциальной.

Другим примером является случай, когда компания AOL выложила в общий доступ предварительно анонимизированный набор данных поисковых запросов для использования в исследовательских целях. Данные пользователей (имена и IP-адреса) были перезаписаны уникальными цифровыми идентификаторами. Однако этого оказалось недостаточно, и на основании сопоставления различных запросов и информации из социальных сетей исследователям набора данных удалось идентифицировать часть пользователей [12, 20].

Компания Netflix для проведения конкурса по повышению эффективности алгоритма прогнозирования рекомендаций фильмов предоставила в открытый доступ набор данных о просмотре фильмов почти полумиллионом пользователей, личные идентификаторы которых были удалены. Однако, как и в случае с AOL, исследователи, сравнив данные Netflix с оценками пользователей фильмов на ресурсе IMDb, обнаружили, что на основе шести оценок фильмов можно идентифицировать пользователя в 84% случаев, а в случае, если известна дата выставления оценки, точность повышается до 99% [12, 20].

Таким образом, обеспечение информационной безопасности DID-систем в настоящее время является актуальной задачей.

2 Текущее состояние обеспечения информационной безопасности информационных систем, реализующих интенсивное использование данных

Особенности текущего состояния обеспечения информационной безопасности (ИБ) информационных систем, реализующих интенсивное использование данных, определяются следующими факторами:

- отсутствие единого общепризнанного определения (эталонной архитектуры) DID-системы – объекта защиты;
- специфичность DID-систем, определяемая их природой (Vs: Variety, Volume, Velocity, Variability);
- относительная молодость DID-систем.

Следствием указанных факторов явилась практическая невозможность построения эффективной системы обеспечения ИБ DID-систем на основе подходов, стандартизированных для традиционных информационных систем. Это привело к необходимости разработки отдельных стандартов и рекомендаций, регламентирующих обеспечение ИБ DID-систем. Таким образом, оценку текущего состояния обеспечения ИБ информационных систем, реализующих интенсивное использование данных, целесообразно делать по следующим направлениям:

- нормативное обеспечение ИБ DID-систем;
- правовое регулирование использования DID-систем и результатов их деятельности.

Состояние по первому направлению может быть охарактеризовано наличием/отсутствием и степенью зрелости международных и/или иных стандартов и рекомендаций, устанавливающих принципы, методы и меры обеспечения ИБ DID-систем.

На текущий момент отсутствуют действующие международные стандарты и рекомендации в области обеспечения информационной безопасности DID-систем. С учетом актуальности проблемы и в отсутствие общепризнанных международных стандартов обеспечения ИБ DID-систем ведущие мировые компании – производители решений построения DID-систем и различные международные и национальные организации разрабатывают собственные подходы к обеспечению ИБ DID-систем [12–19], среди которых внимания заслуживают публикации Cloud Security Alliance (CSA) и Национального института стандартов и технологий США (National Institute of Standards and Technology, NIST) [12–14]. Достоинством этих публикаций является системный подход авторов, включающий:

- разработку эталонной архитектуры DID-систем, то есть объекта защиты (NIST и CSA);
- моделирование угроз безопасности (CSA);

- разработка рекомендаций по применению мер и средств обеспечения ИБ DID-систем (NIST и CSA).

При разработке публикаций в CSA был применен следующий подход, включающий три этапа [12]:

- был проведен опрос членов CSA и специалистов из отраслевых журналов, ориентированных на ИБ, для составления первоначального перечня высокоприоритетных проблем обеспечения ИБ DID-систем;
- были изучены имеющиеся решения для выявленных на первом этапе высокоприоритетных проблем обеспечения ИБ DID-систем;
- высокоприоритетная проблема обеспечения ИБ DID-систем классифицировалась в качестве критической задачи обеспечения ИБ DID-систем, если имеющиеся решения не позволяли в полной мере обеспечить ИБ DID-систем.

По результатам выполнения этих этапов был сформирован перечень из 10 критических задач обеспечения ИБ DID-систем, для которых затем было проведено высокоуровневое моделирование угроз безопасности и разработаны рекомендации в части применения мер и средств обеспечения ИБ [13].

Следует отметить, что уровень зрелости этих публикаций на текущий момент не достаточен для их практического применения с целью построения эффективных систем обеспечения ИБ DID-систем. Так, по мнению специалистов, отвечающих за обеспечение ИБ DID-систем, используемых в Сбербанке, именно недостаточный уровень зрелости документов в совокупности с рядом других факторов привел к необходимости самостоятельно разрабатывать подходы к обеспечению безопасности DID-систем [21].

Второе направление оценки текущего состояния обеспечения ИБ DID-систем связано с правовым регулированием использования как самих DID-систем, так и результатов их деятельности. Основная проблема в области правового регулирования обеспечения ИБ DID-систем заключается в соблюдении требований регуляторных органов в части обеспечения ИБ DID-систем в зависимости от категории доступа обрабатываемой и хранящейся в них информации.

В настоящее время в Российской Федерации отсутствует нормативно-правовая база, регламентирующая использование как самих DID-систем, так и результатов их деятельности. Более того, по мнению некоторых исследователей, основанному на результатах анализа действующих в РФ норм по обеспечению безопасности персональных данных, требования по обеспечению безопасности персональных данных в принципе не могут быть реализованы в DID-системах [20]. С целью решения проблем в области правового регулирования использования как самих DID-систем, так и результатов их деятельности, рабочей группой при администрации Президента в 2016 году начата разработка закона о больших данных, ориентировочные сроки завершения которой намечены на конец 2018 – начало 2019 гг. [22].

Таким образом, текущее состояния обеспечения ИБ информационных систем, реализующих интенсивное использование данных, может быть охарактеризовано следующими особенностями:

- отсутствием действующих общепризнанных международных стандартов в области обеспечения ИБ DID-систем;
- недостаточным уровнем зрелости имеющихся стандартов и рекомендаций для их практического применения с целью построения эффективных систем обеспечения ИБ DID-систем;
- отсутствием в Российской Федерации нормативно-правовой базы, регламентирующей использование как самих DID-систем, так и результатов их деятельности.

3 Направления исследований в области обеспечения ИБ информационных систем, реализующих интенсивное использование данных

Обладая многолетним опытом создания широкомасштабных территориально-распределенных информационных систем, обрабатывающих информацию различных категорий доступа, а также подсистем информационной безопасности для этих систем, авторы считают актуальными следующие направления дальнейших исследований.

1. Разработка принципов и эталонной архитектуры обеспечения информационной безопасности информационных систем, реализующих интенсивное использование данных. Актуальность решения этих задач отмечена, например, в [23–27]. В частности, в [24] рассматривается возможный переход от использования в инфраструктуре доверенных сегментов к принципу «нулевого доверия» (Zero Trust).
2. Исследование информационных систем, реализующих интенсивное использование данных, в части выявления специфичных для них угроз/ уязвимостей. Экосистема DID-систем включает множество компонентов, функционирующих на различных уровнях (инфраструктура, middleware, прикладной). При этом возникают так называемые зависимые уязвимости, при которых уязвимость характерна не для отдельного компонента, а для некоторой их совокупности [28].
3. Исследование возможности и разработка решений обеспечения ИБ информационных систем, реализующих интенсивное использование данных, в соответствии с

требованиями регуляторов в зависимости от категории доступа обрабатываемой в них информации.

Литература

[1] Облака на цифровых вершинах рассеиваются / Коммерсант.ru. www.kommersant.ru/doc/ 3151155

[2] ЦБ перейдет к надзору на основе Big Data / Агентство экономической информации ПРАЙМ. http://1prime.ru/finance/20150917/ 819274417-print.html

[3] Жалобы потребителей финуслуг стали для Центробанка подарком / Известия. http://iz.ru/ news/619529

[4] Big Data в России: оцениваем возможности и риски / CNews. http://www.cnews.ru/reviews/ppt/ 2013_04_05/8.Romanov.pdf

[5] Четвертая парадигма. Научные исследования с использованием больших объемов данных. Под ред. Тони Хея, Стюарта Тэнсли, Кристин Толле / Microsoft Research. https://www. microsoft. com/ru-ru/devcenter/fourthparadigm.aspx

[6] Новые методы работы с большими данными: победные стратегии управления в бизнес-аналитике: Научно-практический сб. Под ред. доктора техн. наук, профессора А.В. Шмида. М.: ПАЛЬМИР, 528 с.: илл. (2016)

[7] Alex Woodie. Why Gartner Dropped Big Data off the Hype Curve / Datanami. https://www. datanami.com/2015/08/26/why-gartner-dropped-big-data-off-the-hype-curve/

[8] Эйден, Эрец: Неизведанная территория: Как «большие данные» помогают раскрывать тайны прошлого и предсказывать будущее нашей культуры / Эрец Эйден и Жан-Батист Мишель; пер. с англ. П. Миронова. М.: Изд-во АСТ, 351 с. (2016)

[9] Big Data: цифровое звено между авиакомпанией и клиентом / Forbes. http://www.forbes.ru/ brandvoice/aeroflot/339961-big-data-cifrovoe-zveno-mezhdu-aviakompaniey-i-klientom

[10] Страшное лицо больших данных / Kaspersky Lab. https://blog.kaspersky.ru/scary-big-data/8676/

[11] How Target Figured Out A Teen Girl Was Pregnant Before Her Father Did / Forbes. https://www.forbes.com/sites/kashmirhill/2012/ 02/16/how-target-figured-out-a-teen-girl-was-pregnant-before-her-father-did/#3d6304f76668

[12] Expanded Top Ten Big Data Security and Privacy Challenges / Cloud Security Alliance, April 2013. https://cloudsecurityalliance.org/download/ expanded-top-ten-big-data-security-and-privacy-challenges

[13] Big Data Security and Privacy Handbook. Cloud Security Alliance, 2016. https:// cloudsecurityalliance.org/download/big-data-security-and-privacy-handbook

[14] NIST Special Publication 1500-4. NIST Big Data Interoperability Framework: Volume 4, Security and Privacy. Final Version 1 / NIST, September 2015. http://dx.doi.org/10.6028/ NIST.SP.1500-4

[15] Big Data Security. Good Practices and Recommendations on the Security of Big Data Systems / ENISA, December 2015. https://www.enisa. europa.eu/publications/big-data-security/at_download/ fullReport

[16] Top Tips for Big Data Security/ IBM, November 2015. https://www.ibm.com/security/whitepapers/ big-data-security-ebook.html

[17] IBM BigInsights Security Implementation: Part 1 Introduction to Security Architecture / IBM, August 2016. http://www.ibm.com/redbooks/ abstracts/ tips1340.html

[18] IBM BigInsights Security Implementation: Part 2 Securing the IBM BigInsights Cluster Perimeter / IBM, December 2016. http://www.ibm.com/ redbooks/abstracts/tips1348.html

[19] Enterprise Security for Big Data Environments / Oracle. http://www.oracle.com/us/technologies/ big-data/big-data-security-wp-3099503.pdf

[20] Савельев, А.И.: Проблемы применения законодательства о персональных данных в эпоху «больших данных» (BIG DATA), Право. Журнал Высшей школы экономики, (1) (2015)

[21] Защита Big Data: проблемы и решения / IT-weekly. http://www.it-weekly.ru/it-news/security/ 117831.html

[22] ФРИИ планирует разработать собственный закон о больших данных / ТАСС. http://tass.ru/ pmef-2017/articles/4308169

[23] ISO/IEC JTC 1, Information Technology, Big Data Preliminary Report 2014 / ISO. https://www.iso. org/files/live/sites/isoorg/files/developing_standards/docs/en/big_data_report-jtc1.pdf

[24] Gualtieri, M., Kindervag, J., MakBig, K.: Data Security Strategies for Hadoop. Enterprise Data Lakes. Apply Zero Trust To Your Big Data Security Strategy. April 25 (2016).

[25] Hrushikesha, Mohanty, Prachet, Bhuyan, Deepak, Chenthati: Studies in Big Data, Volume 11. Big Data, A Primer. Springer India (2015)

[26] Fei Hu: Big Data: Storage, Sharing, and Security. CRC Press (2016)

[27] Zomaya, Albert Y., Sherif Sakr: Handbook of Big Data Technologies. Springer International Publishing AG (2017)

[28] Федорченко, А.В., Чечулин, А.А., Котенко, И.В.: Исследование открытых баз уязвимостей и оценка возможности их применения в системах анализа защищенности компьютерных сетей. Информационно-управляющие системы, (5), сс. 72-79 (2014)

Оценка эффективности систем

System efficiency evaluation

Методика определения интегрального показателя для оценки функционирования центров ЕСИМО

© Е.Д. Вязилов © Н.Н. Михайлов © Д.А. Мельников

Всероссийский научно-исследовательский институт гидрометеорологической информации – Мировой центр данных,
Обнинск, Россия

vjaz@meteo.ru nodc@meteo.ru melnikov@meteo.ru

Аннотация. Представлена методика оценки функционирования распределенных центров данных Единой государственной системы информации об обстановке в Мировом океане (ЕСИМО). Методика включает показатели, отражающие работоспособность аппаратно-программных комплексов центров; актуальность информационных ресурсов, предоставляемых центрами – поставщиками данных; нормативную доступность ресурсов; уровень информационного обслуживания пользователей; обеспечение прав на доступ к ресурсам; обратную связь с пользователями. Приведены примеры оценок функционирования ЕСИМО.

Ключевые слова: центры данных, показатели оценки, работоспособность, актуальность, доступность данных.

Methodology for Evaluating the Functioning of Distributed ESIMO Data Providers

© Evgenii D. Viazilov ©Nick N. Mikhailov © Denis A. Melnikov

All-Russian Research Institute for Hydrometeorological Information – World Data Centre, Obninsk, Russia

vjaz@meteo.ru nodc@meteo.ru melnikov@meteo.ru

Abstract: A methodology for evaluating the functioning of the data providers of the Unified system of information on the situation in the World Ocean is presented. The methodology includes the indicators reflecting the operability of the hardware and software complexes of the centers; the relevance of information resources provided by the centers; normative availability of resources; the level of information service users; Ensuring rights to access resources; help desk from users.

Keywords: data centers, indicators of evaluating, operability, relevance of information, data availability.

Введение

Функционирование Единой государственной системы информации об обстановке в Мировом океане (ЕСИМО) обеспечивается организациями федеральных органов исполнительной власти (ФОИВ): МЧС России, Минобороны России, Минобрнауки России, Минприроды России, Росгидромета, Минпромторга России, Минтранса России, МИД России, Минэнерго России, Росрыболовства, Госкорпорации «Роскосмос» и ФАНО России, назначенными в качестве центров – поставщиков данных в единую систему [3]. Они являются операторами системы и осуществляют ее эксплуатацию на основе Порядков и регламентов деятельности центров (Соглашений о предоставлении информации поставщиками данных в единую систему).

В период постоянной эксплуатации системы проведена оптимизация различных компонентов системы. Так для увеличения скорости обработки данных и уменьшения зависимости от состояния сетей используется кэширование данных с помощью базы интегрированных данных, ресурсы системы обновляются заранее, а не в момент запроса пользователя. Все процессы загрузки данных происходят автоматически. Shape-файлы и картографические сервисы стандарта WMS (Open Geospatial Consortium) строятся автоматически после каждого обновления данных. Для обеспечения актуальности данных применяются режимы планирования обновления данных за счет использования атрибутов метаданных – частоты обновления данных, временного разрешения данных,

Труды XIX Международной конференции «Аналитика и управление данными в областях с интенсивным использованием данных» (DAMDID/ RCDL'2017), Москва, Россия, 10–13 октября 2017 года

указанных в описании каждого ресурса, а также планировщиков обновления данных.

Качество работы операторов системы, связанное с эксплуатацией программного обеспечения (ПО) и поддержкой информационных ресурсов (ИР), отличается. Для оценки их работы вычисляются отдельные показатели – работоспособность аппаратно-программного комплекса (АПК) «Поставщик данных», нормативная доступность ИР и количество обращений к системе. По этим показателям очень трудно дать интегральную оценку работы каждого центра.

В статье описаны методика оценки функционирования распределенных центров данных (рейтинга), алгоритмы расчета показателей и отчеты об оценке функционирования центров ЕСИМО.

1 Методы и средства мониторинга работы информационных систем

Оценка работы государственных информационных систем находит все более широкое применение. Так, еще в 1986 г. был разработан ГОСТ 24.701–86, в котором была определена методика оценки надежности автоматизированных систем управления. Модели оценки надежности автоматизированных систем представлены в работах [2, 8].

В 2002 г. Минкомсвязи России подготовило руководящий документ РД 115.005-2002 [11], в котором предложено организовать мониторинг информатизации России. Задачами этого мониторинга являются:

- сбор первичных данных о состоянии информатизации юридических и физических лиц;
- оценка состояния информатизации физических и юридических лиц;
- анализ результатов мониторинга информатизации Российской Федерации;
- разработка предложений по государственному регулированию процесса информатизации;
- прогноз развития информатизации в результате государственного регулирования;
- мониторинг состояния и развития информатизации в результате реализации государственного регулирования;
- анализ результатов реализации государственного регулирования;
- оценка достоверности результатов мониторинга информатизации.

Наиболее известным результатом использования этого документа является разработка показателей и расчет рейтинга сайтов ФОИВ, высших исполнительных органов государственной власти субъектов РФ и администраций муниципальных образований.

В последние годы стали развиваться методы оценки работы распределенных информационных систем [1, 4, 9] и открытых данных [6]. Ярким примером оценки работы распределенной системы являются сведения о работе портала «Госуслуги» [10].

Для организации мониторинга работы портала панъевропейского проекта Sea Data Net [12], в рамках которого имеется более 100 поставщиков данных, используется система мониторинга компьютерных систем и сетей Nagios. В этой системе оцениваются отдельные показатели: надежность работы провайдеров данных; количество зарегистрированных пользователей; выполненных запросов; количество введенных в базу метаданных экземпляров сведений о массивах данных, организациях, проектах, рейсах, платформах; поддерживаемых провайдерами данных на своих серверах записей океанографических данных.

При выборе наиболее экологически благоприятных мест также можно использовать оценку рейтингов мест проживания [4].

Наиболее продвинутая методика для оценки органов государственной власти при составлении рейтинга публикации информации в форматах открытых данных представлена в [7]. В этой методике выделены три группы показателей:

- качество размещенных наборов данных, включающее актуальность опубликованных данных, корректность метаданных, пользовательскую оценку наборов данных;
- востребованность наборов данных – количество скачиваний;
- выполнение требований законодательства Российской Федерации (распоряжений Правительства, Федеральных органов исполнительной власти, органов исполнительных власти субъектов РФ, планов публикаций).

Для оценки работы ЕСИМО используется система мониторинга ресурсов и сервисов (МРС). В этой компоненте заложены метрики, которые содействуют пониманию того, достиг ли центр требуемых значений показателей работы, например, уменьшения времени загрузки данных в базу данных до 60 мин. для самых объемных ресурсов, или повышения актуальности интегрированных данных (обновление данных происходит в соответствии с установленным в метаданных регламентом), или увеличения надежности работы АПК до 96.5%. При мониторинге работы системы для различных метрик используются пороговые или референтные значения, а также метод базовых линий.

Пороговые значения установлены в процентах для надежности работы АПК, доступности ИР. Для некоторых значимых метрик используются референтные значения, содержащиеся в рекомендациях производителей программного обеспечения и аппаратных средств, например, это касается наличия свободной оперативной памяти (не меньше 10%) или памяти на диске (не меньше 30%).

Метод базовых линий характеризует нормальную работу системы. Показатель получается путем обработки предыдущих результатов оценки, например, использования ИР. На основе полученных результатов выбираются три категории ИР (активно используемые, слабо используемые и плохо используемые). На сегодняшний день к первой категории относятся ресурсы, которые загружались

больше 100 раз в месяц, ко второй – от 5 до 100, к третьей – меньше 5.

Эффективность любой информационной системы существенно зависит от правильности выбора пороговых значений метрик, характеризующих качество ИТ-сервисов. Соглашение об уровне сервиса – SLA (Service Level Agreements), составленное между поставщиком данных и технологическим центром по поддержке системы, включает именно такие показатели работы системы. В соглашении указывается компонент «Обратная связь», с помощью которого пользователь может обратиться в службу поддержки путем заполнения специальной формы с претензией.

Если ПО работает хорошо, а актуальность данных недостаточна (данные старые), то пользователь будет недоволен. Для измерения доступности ИР недостаточно пинговать оборудование и ПО, нужно определить, при каких значениях и каких метриках сервис работает хорошо, а при каких – не очень. В идеале система должна быть настроена таким образом, чтобы обеспечивалась ее способность к самовосстановлению в случае различных сбоев, аварий, выхода из строя отдельных узлов АПК.

Недостатками применяемых методик расчета показателей работы автоматизированных систем является отсутствие средств сравнения работы поставщиков данных.

2 Система мониторинга ресурсов и сервисов ЕСИМО

В соответствии с «Руководством по функционированию ЕСИМО» показателями работы единой системы являются техническая доступность (работоспособность) АПК центров; актуальность и доступность ИР, предоставляемых центрами данных; нормативная доступность ИР; посещаемость порталов. Чтобы повысить надежность работы инфраструктуры ЕСИМО и обеспечить координацию работы службы технической поддержки используются компоненты: МРС, включающий контроль работы АПК и доступности ИР; «Обратная связь», которая дает возможность подать замечания, предложения, сведения об ошибках; «Отчетность и статистика» – подготовка регулярных отчетов. Информация о работе ЕСИМО необходима как для внешней отчетности, так и для решения тактических вопросов управления системой.

Компонент МРС построен на основе инструмента Zabbix и реализует мониторинг АПК системы, работу сетей. Если раньше основными типами ошибок, влияющими на надежность системы, были сбои инструментальных средств (переполнение оперативной памяти при работе сервера приложений JBoss), аппаратных средств (отсутствие резервирования), то сейчас сбои происходят в основном на сетевом уровне. В случае резкого повышения числа запросов, система может не справиться с ними из-за отсутствия балансировки нагрузки.

Если базе интегрированных данных не хватает производительности, то загрузка ИР выстраивается в очередь. Для гарантированной доставки ресурсов запрос на доставку стоит в очереди, пока не будет выполнен. Для инициализации ПО на различных узлах системы созданы образы виртуальных машин, с помощью которых достаточно быстро перезагружается любой компонент. При работе компонентов иногда возникают ситуации, когда необходима перезагрузка всей системы (отключение электроэнергии, выход из строя сервера, операционной системы). Последовательность запуска компонентов и соответствующие задания заранее оформлены и запускаются автоматически.

Недостатком МРС является то, что инженер службы технической поддержки слишком поздно узнает о неисправности, он видит ситуацию, когда инцидент уже состоялся. Поэтому необходимо применение проактивной технологии мониторинга работы системы, когда прогнозируется возможный сбой до его наступления. Многое, что ранее было сферой ответственности оператора или администратора системы, теперь автоматически выполняет МРС: в реальном времени отслеживаются недостаток дискового пространства, использование процессора и оперативной памяти, проверяется статус серверов. Если в сети назревает проблема, то срабатывают предупреждения и появляется информация об инциденте. Если инцидент невозможно разрешить автоматически, на помощь приходят средства удаленного доступа к серверам.

Кроме оперативного применения результатов МРС необходимо отслеживать и долговременные тенденции изменения различных показателей, сравнить работу всех центров. И здесь уже отдельных показателей недостаточно, необходимо проведение интегральной оценки работы центров системы.

3 Система показателей функционирования ЕСИМО

3.1 Группы показателей

При разработке методики интегральной оценки функционирования центров данных ЕСИМО использован подход, представленный в «Методике оценки органов государственной власти при составлении рейтинга публикации информации в формате открытых данных» [7]. Целью разработки этой методики является сравнение работы центров данных, чтобы на основе интегральной оценки можно было по количеству полученных баллов определить, как они работают. Такая оценка позволяет центрам данных определить слабые места в своей работе. Для интегральной оценки функционирования определены следующие группы показателей:

- работоспособность АПК центров;
- актуальность ИР, предоставляемых центрами;
- нормативная доступность ИР;
- уровень информационного обслуживания пользователей;
- обеспечение прав на доступ к ИР, отнесенным к информации, предоставляемой на условиях

центра данных;
- обратная связь с пользователями системы.

Наиболее важными группами показателей с точки зрения пользователей являются работоспособность АПК, актуальность ИР и уровень информационного обслуживания, поэтому этим показателям определен наибольший вклад в интегральную оценку работы центров единой системы, таблица 1. В дальнейшем, когда работоспособность АПК и актуальность ИР достигнут плановых значений, их вклад будет уменьшен, а будет увеличен вклад группы показателей по информационному обслуживанию пользователей. При реализации этой методики на программном уровне эти вклады можно будет устанавливать при настройке приложения.

Таблица 1 Вклад каждой группы показателей

Наименование группы показателей	Обозначение	Вклад
Работоспособность АПК	К1	0.3
Актуальность ИР	К2	0.3
Нормативная доступность ИР	К3	0.05
Уровень информационного обслуживания пользователей	К4	0.25
Обеспечение прав на доступ к ИР	К5	0.05
Обратная связь с пользователями	К6	0.05

Работоспособность АПК

Показателями работоспособности АПК центров данных в ЕСИМО являются: работоспособность АПК; время простоя АПК (в часах).

Работоспособность (или надежность) АПК (**$К_1$**) – это относительная величина, характеризующая процент времени, когда ПО «Поставщик данных» и другие компоненты АПК центра данных были работоспособны. При этом необходимо исключать из расчетов период времени, затрачиваемый на профилактические работы (**Тпроф**). Общее время работы Поставщика данных (**То**) за отчетный период – это количество времени в часах (**Кч**) за отчетный период минус время профилактических работ, оно равно: **То=Кч–Тпроф**. Например, за сентябрь То=30*24–0=720 часов.

Время простоя (**Тп**) – это период времени, когда ПО «Поставщик данных» или другой компонент системы не работал (был не доступен). Информация о времени простоя берется из компонента МРС и готовится на основе следующего перечня объектов мониторинга и снимаемых значений метрик по проблемным событиям для виртуальных машин (закончилась свободная оперативная память, на диске закончилось свободное место, процесс JBoss исчерпал отведенную память); сетевая доступность (транспортная доступность административного интерфейса – закрыт порт 8081, недоступно сетевое соединение); прикладное приложение (не работает веб-интерфейс JBoss, нет дочерних Java-процессов, остановился процесс JBoss, не работает административный интерфейс ПО «Поставщик данных», нет доступности веб-сервиса). Каждая метрика имеет значение триггерной функции, равное единице, когда значение метрики находится в рабочих пределах, и нулю – в аварийных значениях. Если хотя бы одна из метрик имеет значение нуль, то вся система помечается как нерабочая. Время неработоспособности равно периоду времени существования проблемы на любом из компонентов поставщика данных. Общая сумма времени простоя складывается из выявленных диапазонов времени простоя. Компонентами, по которым ведется мониторинг на центральном, региональных и специализированных узлах, являются портал, геоинформационная система, сервисная шина, сервер интеграции и база интегрированных данных. Для центров системы компонент, по которому ведется мониторинг, – это Поставщик данных. По каждому компоненту ведется контроль следующих объектов: виртуальная машина, сеть, прикладное приложение. Для каждого объекта осуществляется контроль его составных частей и проблем, фиксированных для каждой части. В случае сетевой недоступности узла будет фиксироваться только проблема сетевой недоступности, поскольку остальные проблемы физически не могут быть зафиксированы. Работоспособность **$К_1$=(То–Тп)/То**.

3.3 Актуальность информационных ресурсов

Показателями ИР, предоставляемых центрами, являются их количество и актуальность**.**

Количество ИР (N) – это число описанных в виде структурированных файлов, или таблиц баз данных, или объектных файлов данных центра, представленных ими в ИР системы с помощью ПО «Поставщик данных». Этот показатель оценивается на основе ежедневной автоматизированной проверки количества единиц ресурсов и готовности источников данных предоставить информацию. Количество поддерживаемых ресурсов в конкретный момент времени можно определить в каталоге ИР по адресу http://esimo.ru/portal/portal/esimo-user/data и в административном разделе портала по справке «Показатели ИР» по адресу http://esimo.ru/portal/portal/ admin/stat/stat-provider.

Актуальность ИР ($К_2$**)** – это соответствие обязательств центра по обновлению данных, объявленных в описании ресурса (атрибут «Частота обновления»), реальным дате и времени обновления в системе. В ЕСИМО все ИР обновляются от 10 мин. до одного раза в год. Для структурированных данных и таблиц баз данных этот показатель оценивается на основе ежесуточной автоматической проверки времени обновления ИР в сопоставлении со значением этого показателя, указанным (заявленным) при регистрации ресурса. Для объектных файлов актуальность оценивается на основе даты изменения файла, формируемой операционной системой или отражаемой в имени файла в соответствии с принятыми правилами именования файлов. Также необходима проверка времени обновления метаданных в случаях, если данные не содержат параметр «Дата и время» либо не включен или неправильно настроен планировщик актуализации. В результате проверки актуальности ИР ежедневно выявляется число актуальных

ресурсов (N_a). В дальнейшем предполагается оценивать этот показатель не только с точки зрения невыполнения регламента обновления данных, но и с точки зрения реального времени их опоздания. На данном этапе развития системы и отсутствия финансирования для поддержки ИР ужесточать требования по актуальности (учитывать время отставания по обновлению ИР) нецелесообразно.

Показатель актуальности (**$К_2$**) – это удельный вес актуальных ресурсов в общем количестве ресурсов центра, он вычисляется как среднее отношение числа штатно обновляемых ресурсов к общему числу ресурсов за отчетный период по формуле **$К_2=N_a/N$.**

3.4 Нормативная доступность информационных ресурсов

Нормативная доступность устанавливается обладателями информации в Порядках и регламентах деятельности центров ЕСИМО путем присвоения информации одной из категорий: «свободно распространяемая информация» (или открытая) и «информация, предоставляемая по договору – соглашению с обладателем информации».

Нормативная доступность (**$К_3$**) оценивается как отношение числа ИР со «свободно распространяемой информацией» (**Nд**) к общему числу ресурсов на последний день отчетного периода: **$К_3=Nд/N$.**

3.5 Уровень информационного обслуживания пользователей

При расчете показателя «Уровень информационного обслуживания пользователей ЕСИМО» (**$К_4$**) учитываются составляющие:

- количество обращений всех категорий пользователей (единиц) к ИР центра (**$К_{41}$**);
- число загрузок (единиц) ИР для просмотра или скачивания (**$К_{42}$**);
- число загрузок геосервисов (просмотров слоев), подготовленных по информации центра (**$К_{43}$**);
- количество доставок ИР по подписке ($К_{44}$);
- востребованность автоматизированных рабочих мест (АРМ) пользователей, находящихся в ведении центра (по числу обращений пользователей) (**$К_{45}$**).

Количество обращений всех категорий пользователей (единиц) к ИР центра определяется за отчетный период по формуле **$К_{41}=К_{обр41}/К_{ср41}$.** Для оценки вклада центра применяется нормирование на среднее значение того или иного показателя (**$К_{ср}$**) в целом для ЕСИМО**.**

Число загрузок ИР для просмотра или скачивания вычисляется на основе числа обращений к таблицам базы интегрированных данных с ресурсами центров по формуле **$К_{42}=К_{обр42}/К_{ср42}$**.

Число загрузок геосервисов (просмотров слоев) определяется на основе числа обращений к URL-адресу сервиса из ГИС-вьювера OceanView, порталов, АРМов и других приложений: **$К_{43}=К_{обр43}/К_{ср43}$**.

Количество доставок ИР по подписке за выделенный период – это число востребованных ресурсов каждого центра: $К_{44}=К_{обр44}/К_{ср44}$. Если доставок нет, этот показатель не вычисляется.

Востребованность АРМов пользователей, находящихся в ведении центра, оценивается по числу обращений пользователей **$К_{45}=К_{арм}/К_{армср}$**. Если АРМов нет, то показатель не вычисляется.

Показатель уровня информационного обслуживания пользователей равен сумме его составляющих: **$К_4=К_{обр41}+К_{обр42}+К_{обр43}+К_{обр44}+К_{45}$.**

3.6 Обеспечение прав на доступ к информационным ресурсам

Чтобы получить доступ к закрытым ИР, надо получить разрешение их обладателя посредством запроса в соответствующий центр. Предлагается ввести показатель обеспечения прав на доступ к ИР (**$К_5$**) и вычислять его по формуле **$К_5=К_{нвып5}/К_{общ5}$.** Характеристиками обеспечения прав на доступ к ИР являются общее число запросов на получение разрешения **($К_{общ5}$)** и число невыполненных запросов (**$К_{нвып5}$**) на доступ к ИР.

3.7 Обратная связь с пользователями

Характеристиками функционирования ЕСИМО являются также число запросов, полученных через компонент «Обратная связь» с количеством установленных фактов ненадлежащего информационного обслуживания (**$Ф_{нн6}$**) и обоснованных замечаний относительно действий (бездействия) администраторов узлов (**$Ф_{оз6}$**). Удельный вес ресурсов, для которых была высказана негативная пользовательская оценка в общем количестве ресурсов (**$К_6$**), вычисляется по формуле **$К_6=Ф_{оз6}/Ф_{нн6}$.**

3.8 Расчет интегральной оценки

Интегральная оценка центра (**$Р_{центра}$**) ЕСИМО рассчитывается путем вычисления отношений значений каждого показателя (кроме **$К_4$**) к среднему значению этого показателя в целом для ЕСИМО по следующей формуле с учетом веса каждой группы показателей: **$Р_{центра}=0.3К_1/К_{1ср}+0.3К_2/К_{2ср}+0.05К_3/К_{3ср}+$ $+0.25К_4+0.05К_5/К_{5ср}+0.05К_6/К_{6ср}$.**

4 Формирование отчетов об оценке функционирования центров

Отчет с показателями работы центров ЕСИМО размещен на портале в виде автоматически обновляемой страницы (http://portal.esimo.ru/portal/ stat/), а также включен в качестве дополнительных разделов в общий отчет о функционировании системы, подготавливаемый ежеквартально, рассылаемый в центры системы и предоставляемый для рассмотрения в Межведомственную комиссию по ЕСИМО.

ПО компонента «Статистика и отчетность» обеспечивает автоматизированную подготовку и поддержку актуальности показателей, группируемых в виде:

- оценки функционирования ЕСИМО (таблица 2);
- интегральной оценки работы центров ЕСИМО – рейтинг центров (таблица 3);

- оценки центров по показателю «Количество ресурсов и их востребованность» (таблица 4);
- оценки качества ИР и негативных оценок этих ресурсов.

Таблица 2 Общая статистика функционирования ЕСИМО в 2015–2016 гг.

Показатели	Значение	
	2015	2016
Количество центров ЕСИМО, ед.	37	37
Работоспособность АПК, в %	90,6	93,8
Общее количество ресурсов, размещенных на портале, ед.	3211	3444
Количество подготовленных новых ресурсов, ед.	20	20
Актуальность информационных ресурсов, в %	87,4	88,8
Кол-во актуальных информационных ресурсов на момент отчета, ед.	-	3057
Нормативная доступность информационных ресурсов, в %	56,6	72,1
Общее количество просмотров (скачиваний) ресурсов, ед.	130673	219597
Среднее количество скачиваний ресурсов, ед.	6534	10902
Удельный вес просмотренных (скачанных) ресурсов, в общем количестве ресурсов	0,59	0,51
Общее количество ресурсов, переданных по подписке, ед.	-	169041
Обеспечение прав на доступ к ИР ЕСИМО, количество выданных ролей на доступ, ед.	-	758
Обратная связь с пользователями ЕСИМО, кол-во рекламаций, ед.	-	-

Кроме рейтинга центров ЕСИМО компонент «Статистика и отчетность» обеспечивает расчет показателей в абсолютных значениях:

- сведений о работоспособности АПК центров данных, таблица 5;
- сведений о показателях ИР – актуальность и нормативная доступность ресурсов, таблица 6;
- статистики обращений к порталу, таблица 7;
- востребованности АРМов пользователей;
- количества доставок ИР по подписке;
- выполнения запросов на роли для доступа к ИР системы;
- количества пользователей и полученных ролей.

Выводы

Разработана методика определения интегрального показателя по оценке функционирования центров ЕСИМО в виде рейтинга центров, которая, с одной стороны, расширяет принятый ранее состав показателей работы с целью детализации оценки работы единой системы, а, с другой стороны, позволяет выделить хорошо и плохо работающие центры. Впервые в этой работе предложен комплекс показателей для оценки функционирования центров ЕСИМО, некоторые из которых ранее рассматривались только отдельно или вообще не использовались в других системах (например, нормативная доступность и обеспечение прав на доступ к ИР). Модель позволяет сравнить любое количество центров системы. В настоящее время оценивается 28 из 37 центров (там, где установлены программные агенты компонента МРС).

Показатели работы центров включаются в ежеквартальные Отчеты о функционировании системы. Отчеты регулярно рассматриваются на заседаниях Межведомственной комиссии по ЕСИМО, при этом отмечаются плохо работающие центры и принимаются соответствующие решения в адрес ФОИВ, в ведении которых находятся эти центры. Использование результатов оценки функционирования центров позволило увеличить показатели их работы – работоспособность до 96.5%, актуальность – до 90%. В 2013 г. надежность работы ЕСИМО составляла 87%. Большинство центров уже сейчас имеет показатель работоспособности – 99.9%, а актуальность ИР – до 95%. Введение показателя «Нормативная доступность» позволило увеличить число открытых ИР с 50% до 70%.

Таблица 3 Рейтинг центров данных в 2016 г.

Центр ЕСИМО	Работо-способность	Актуаль-ность	Норм. доступ	Информационное обслуживание						Права на доступ	Обратная связь	Баллы	Место
ФГБУ «ВНИИГМИ-МЦД»	0,31	0,3	0,05	2,66	2,26	4.08	4,92	2,26	16.18	0	0	16,84	1
ФГБУ «Гидрометцентр России»	0,29	0,29	0,07	0,73	2,09	0,14	0	0,17	3.13	0	0	3,78	2
ФГУП «ЦНИИ «Центр»»	0,32	0,33	0,07	0,14	0	0	0	0,45	0.59	0,01	0	1,32	3
ФГБУ «ДВНИГМИ»	0.32	0,32	0,07	0,32	0	0,09	0	0,18	0.59	0	0	1,3	4
ФГБУ «ААНИИ»	0,23	0,22	0,06	0,23	0	0,15	0	0,32	0.7	0,04	0	1,25	5
ФГУП «Морсвязьспутник»	0,32	0,29	0,07	0,01	0,07	0,21	0,01	0,26	0.56	0	0	1,24	6
ФКУ НЦУКС	0,32	0,3	0,0	0,03	0	0,03	0,01	0,45	0.52	0	0	1,14	7
ФГБУ ЦСМС	0,32	0,34	0,0	0,01	0,29	0,02	0	0,15	0.47	0	0	1,13	8

Таблица 4 Количество ресурсов и их востребованность в 2016 г.

Центр ЕСИМО	Кол-во ИР	Кол-во скачиваний	Уд. вес скачанных ИР
ФГБУ «ВНИИГМИ-МЦД»	1620	115945	0,41
ФГБУ «Гидрометцентр России»	103	31919	0,65
ФГУП «ЦНИИ «Центр»	12	6225	1.0
ФГБУ «ДВНИГМИ»	334	13820	0,65
ФГБУ «ААНИИ»	284	10128	0,3
ФГУП «Морсвязьспутник»	18	274	0,94
ФКУ НЦУКС	22	1387	1.0
ФГБУ ЦСМС	20	148	0,64
НЦ ОМЗ ОАО «РКС»	69	4047	0,6
……...	….	….	…...
В целом ЕСИМО	**2883**	**218048**	**0,66**

Таблица 5 Показатели работоспособности АПК узлов ЕСИМО в 2016 г.

Центр ЕСИМО	Работоспо-собность,%	Время простоя,ч
ФГБУ «ВНИИГМИ-МЦД»	99,84	3,59
ФГБУ «ДВНИГМИ»	98,33	36,41
ФГБУ «ААНИИ»	93,36	145,06
……..		
В целом по ЕСИМО	96.5	500

Таблица 6 Показатели информационных ресурсов ЕСИМО 2016 г.

Центр ЕСИМО	Информационные ресурсы		
	Всего ед.	Актуаль-ность,%	Норм. доступ.,%
ФГБУ «ВНИИГМИ-МЦД»	1848	67,29	66
ФГБУ «ДВНИГМИ»	334	92,10	99
ФГБУ «НИЦ «Планета»	51	46,92	92
……			
В целом по ЕСИМО	3500	90	70

Таблица 7 Статистика обращений к порталу ЕСИМО в 2016 г.

Год	Посещений ед.	Пользователей, ед.	
		Уникальных	Зарегистрированных
2016-07	118464	4005	36
2016-08	170727	4619	50
2016-09	164579	5015	50

Литература

[1] Акимова, Г.П., Соловьев, А.В.: Методология оценки надежности иерархических информационных систем. Труды ИСА РАН, 23, сс. 18-47 (2006)

[2] Василенко, Н.В., Макаров, В.А.: Модели оценки надежности программного обеспечения. Вестник Новгородского гос. ун-та, (28), сс. 126-132 (2004)

[3] Вязилов, Е.Д., Михайлов, Н.Н.: Интеграция данных о морской среде и деятельности. Инфраструктура спутниковых геоинформационных ресурсов и их интеграция. Сб. науч. статей под ред. М.А. Попова и Е.Б. Кудашева. Киев: Карбон-Сервис, сс. 174-181 (2013)

[4] Коробов, В.Б.: Экспертные методы в географии и геоэкологии. Архангельск: Поморский университет, 236 с. (2008)

[5] Методика мониторинга и оценки востребованности открытых данных. М.: Аналитический центр при Правительстве РФ, 44 с. (2013)

[6] Методика оценки органов государственной власти при составлении рейтинга публикации информации в формате открытых данных (проект). М.: Аналитический центр при Правительстве РФ, 20 с. http://ac.gov.ru/files/attachment/7856.pdf (2016)

[7] Методические рекомендации по публикации открытых данных государственными органами и органами местного самоуправления, а также технические требования к публикации открытых данных. Версия 3.0, 101 с. (2014)

[8] Громов, Ю.Ю. и др.: Надежность информационных систем. Тамбов: Тамбовский гос. техн. ун-т, 160 с. (2010)

[9] Павский, В.А., Павский, К.В.: Стохастическое моделирование и оценка размера структурной избыточности масштабируемых распределенных вычислительных систем. Изв. ЮФУ. Технические науки, (12 (161)), сс. 66-73 (2014)

[10] Подведены итоги работы портала госуслуг в первом полугодии 2016 года. Минкомсвязь России. http://minsvyaz.ru/ru/events/35420 (2016)

[11] РД 115.005-2002. Информационные технологии. Мониторинг информатизации России. Основные положения мониторинга. Министерство РФ по связи и информатизации. Утв. 4 марта 2002 г. 54 с. Утвержден информационным письмом Минсвязи России от 4 марта 2002 г. № 1341. Дата введения в действие – 4 марта 2002 года

[12] Pan-European infrastructure for ocean & marine data management. Project Sea Data Net. https://www.seadatanet.org/

Имитационное моделирование данных для определения готовности муниципальных образований к внедрению технологий Smart City

© О.О. Комаревцева

Среднерусский институт управления – филиал Российской академии народного хозяйства и государственной службы при Президенте Российской Федерации,
Орел, Россия

komare_91@mail.ru

Аннотация. Цель исследования заключалась в определении степени готовности муниципальных образований Российской Федерации к внедрению технологий Smart City. Предложена имитационная модель, позволяющая определить степень готовности муниципальных образований к внедрению технологий Smart City, подобрать городские проекты (Smart-проекты), наиболее релевантные существующему уровню готовности, выявить основные барьеры на пути их реализации. В ходе исследования использованы методы структурного и графического анализа, суммарной оценки и рейтингов, группового учета аргументов. Область применения полученных результатов достаточно обширна. Прежде всего, данное исследование будет интересно ученым, занимающимся разработками в области цифровой экономики, управления данными Big data, а также практическим специалистам, реализующим проекты эффективной урбанизации городской среды.

Ключевые слова: Smart City, имитационное моделирование, интенсивное использование данных, технологии, цифровая экономика.

Simulation of Data for Determining the Readiness of Municipalities to Implement Smart City Technologies

© O.O. Komarevtseva

Srednerusskiy Institute of Management-branch of Russian Academy of National Economy and Public Administration,
Orel, Russia

komare_91@mail.ru

Abstract. The purpose of the study is to determine the degree of readiness of municipal municipalities in the Russian Federation to implement Smart City technologies. The author offers an imitation model that allows to determine the degree of readiness of cities to implement Smart City technologies, to select Smart projects, to identify the main barriers to the implementation of Smart projects. In the course of the research methods of structural and graphical analysis, summary evaluation and ratings, group accounting of arguments were used. The field of application of the results obtained is quite extensive. This study will be interesting to scientists engaged in developments in the field of digital economy, data management Big data, practical specialists implementing urban urbanization projects.

Keywords: Smart City, simulation data intensive domains, technology, digital economy.

1 Введение

Экономическая парадигма XXI века претерпевает изменения, связанные с переходом к цифровому развитию федеральной и региональной экономики. Кроме того, в нынешних условиях жесткой конкуренции городов как внутри страны, так и за ее пределами, вопрос внедрения технологий цифровой экономики становится особенно актуальным. Стремительно развиваясь, муниципальные образования формируют новые экономические и культурные центры, которые впоследствии стимулируют экономические изменения. При этом депрессивные муниципальные образования испытывают в данном процессе некоторые проблемы. К их числу относятся: неравенство муниципальных образований Российской Федерации в отношении доступа к цифровым системам; необеспеченность социальной вовлеченности органов местного самоуправления в данный процесс; депрессивно-стагнирующее состояние экономик некоторых городов; отсутствие фундаментальных

Труды XIX Международной конференции «Аналитика и управление данными в областях с интенсивным использованием данных» (DAMDID/ RCDL'2017), Москва, Россия, 10–13 октября 2017 года

основ построения цифровой экономики. В то же время развитие экономики муниципального образования невозможно без привлечения инвестиций, обеспечения потребностей населения в новых интеллектуальных услугах, эффективного управления городской инфраструктурой. Для реализации данных направлений в современных условиях развития требуется внедрение элементов интеллектуальной экономики, одними из которых выступают технологии Smart City. Однако возникают вопросы: готовы ли муниципальные образования к внедрению технологий Smart City? На основе какой методики или модели должна быть осуществлена оценка готовности?

В соответствии с выделенной проблематикой целью исследования выступает определение степени готовности муниципальных образований Российской Федерации к внедрению технологий Smart City. Для реализации поставленной цели необходимо решить следующие задачи:

- рассмотреть теоретико-методическую часть исследования вопроса имитационного моделирования данных как инструмента, определяющего уровень готовности муниципальных образований к внедрению технологий Smart City;
- выявить основные рейтинги оценки степени соответствия городов принципу Smart City, используемые в современных реалиях развития территорий;
- применить имитационное моделирование для определения готовности муниципальных образований к внедрению технологий Smart City.

2 Теоретико-методические аспекты готовности муниципальных образований к внедрению технологий Smart City

Исследование теоретико-методического инструментария готовности муниципальных образований к внедрению технологий Smart City прослеживается в трудах зарубежных и российских ученных. Отметим, что многие из исследований основаны на определении готовности городов к внедрению технологий Smart City через построение статистической модели. Ключевыми моделями этих исследований выступают: регрессионная статистическая модель экспериментальных данных (определяющая закономерности между экономическим положением муниципального образования и инновационным развитием городской среды) [7], модель статистических испытаний (основанная на многократном теоретико-вероятностном и статистическом моделировании параметрических величин концепции Smart City) [10], статистическая модель робастности полученных результатов (позволяющая определить устойчивость / неустойчивость развития муниципального образования с учетом изменений параметров случайных величин и начальных условий моделирования) [1], статистическая модель матрицы переходов (устанавливающая параметры масштабирования и перемещения для изменения вектора развития муниципального образования) [4], модель оценки сомнительных результатов (выявляющая наличие факторов, не оказывающих существенного влияния на достижение запланированного результата в рамках развития муниципального образования) [8]. Однако в условиях анализа большого количества данных требуются автоматизация предлагаемых показателей (индикаторов) оценки или применение инструментов имитационного моделирования. Имитационное моделирование данных является одним из удобных и практических инструментов определения готовности муниципальных образований к внедрению технологий Smart City. В рамках представленного выше инструмента исследования интересными представляются следующие модели: агентная модель (используется для определения индивидуальных свойств и правил поведения активных агентов в процессе внедрения концепции Smart City в городскую среду) [9], модель системной динамики (применяется для выявления существенных характеристик объектов, явлений, процессов, в рамках концепции Smart City, с последующем установлением причинно-следственных связей между данными категориями) [3], модель детерминации (трансформируется в соответствии с изменениями внешней среды и адаптируется под создаваемые условия развития муниципального образования) [2], дискретно-событийная модель (акцентирует внимание на ключевых процессах экономического развития и абстрагируется от непрерывных событий, происходящих в муниципальном образовании) [11].

Несмотря на большое количество исследований, проводимых на основе названного инструментария, к сожалению, отсутствуют единая методика или российский стандарт оценки готовности муниципальных образований к внедрению технологий Smart City. Международные модели и рейтинги исследования готовности городов к внедрению технологий Smart City находятся на стадии апробации и используются для решения конкретных практических вопросов: определение устойчивости города, наличие в городской среде элементов «умного города», выявление показателей интеллектуализации городской инфраструктуры и т. д. Этот аспект подтверждает актуальность темы и обосновывает новизну исследования, проявляемую в авторском подходе к имитационному моделированию данных для определения степени готовности муниципальных образований к внедрению технологий Smart City.

Рассмотрим ключевые рейтинги оценки, применяемые для определения готовности муниципальных образований к внедрению технологий Smart City.

3 Рейтинги оценки степени соответствия городов принципам концепции Smart City

Для оценки степени готовности муниципальных

образований выбраны три ключевых рейтинга соответствия городов принципам концепции Smart City (таблица 1).

Таблица 1. Ключевые рейтинги соответствия городов принципам концепции Smart City

Название рейтингов	Индикаторы
Рейтинг соответствия городов принципам концепции Smart City	Умная экономика (smart economy), умная мобильность (smart mobility), умный подход к окружающей среде (smart environment), умные люди (smart people), умный образ жизни (smart living), умное правительство (smart governance)
Рейтинг устойчивого развития городов Российской Федерации	Экономика, городское хозяйство, социальная сфера, экологическая обстановка
Система показателей умных городов	Экономика (ИКТ, инновации, занятость, торговля, производительность, физическая инфраструктура), окружающая среда (качество воздуха, водоснабжение, шум, биоразнообразие, энергетика), общество и культура (образование, здравоохранение, безопасность, жилье, культура, социальная вовлеченность)

Одноименный рейтинг соответствия городов принципам концепции Smart City разработан лабораторией Венского технического университета[5]. Основой этого рейтинга является рассмотрение европейских городов (с численностью населения до 1 млн. человек) на предмет соответствия принципам концепции Smart City. Названный рейтинг включает два ключевых блока с шестью характеристиками «умного города». Рассмотрим их более подробно.

Блок 1. Открытость и способность социальных институтов к быстрой трансформации и модернизации:

- «умная экономика» (smart economy): инновационное развитие, уровень развития предпринимательства, гибкость рынка труда; включенность в международное экономическое пространство, экономический образ города, экономическая продуктивность – индикаторы не устойчивы, изменяются в зависимости от ситуации; рассчитываются в процентном соотношении;

- «умный подход к окружающей среде» (smart environment): уровень устойчивого управления ресурсами, степень загрязненности воздуха, уровень обеспокоенности экологической средой – измерение влияния технологического прогресса на уровень экологии;

- «умная мобильность» (smart mobility): инновационная и безопасная транспортная система, возможность без проблем добраться во все районы и места города, открытость города на национальном и интернациональном уровнях, доступность информационно-коммуникационных технологий в городской инфраструктуре – наличие высокотехнологической базы совместно с информационной доступностью.

Блок 2. Уровень образованности и социальной активности населения:

- «умные люди» (smart people): степень образованности горожан, уровень квалификации населения, способность и желание обучаться на протяжении всей жизни, социальное и этническое многообразие в разрезе городского населения – важным является определение открытости горожан к новым изменениям;

- «умный образ жизни» (smart living): уровень здоровья населения, уровень индивидуальной безопасности граждан, туристическая привлекательность города, социальная сплоченность населения, качество проживания и уровень развития жилищно-коммунальной системы, доступность образовательных учреждений, развитость инфраструктуры культурного пространства – определение степени участия граждан в принятии решений по модернизации городского пространства;

-«умное правительство» (smart governance): участие городского населения в принятии решений в области развития города, уровень работы общественных и социальных организаций, прозрачность работы институтов управления – определение наличия компонентов умного управления.

Представленный выше рейтинг обладает качественными преимуществами, связанными с подробным исследованием соответствия городов принципам концепции Smart City (минимально – 74 показателя). Однако в некоторых случаях эти преимущества выступают как недостаток, например, когда требуется простая методика определения готовности внедрения технологий Smart City для переговоров потенциальных инвесторов и органов местного самоуправления.

Вторым выступает рейтинг устойчивого развития городов Российской Федерации, сформированный агентством “Sustainable growth management agency” (ООО «Агентство ЭС ДЖИ ЭМ»)[6]. В основу этого рейтинга включены тридцать статистических показателей, характеризующих муниципальное образование по следующим критериям: состояние экономики, городского хозяйства, социальной сферы, а также экологической обстановки. Объектами выборки выступают города – административные центры субъектов Российской Федерации. Особенностью данного рейтинга является тот факт, что высокие параметрические показатели не всегда

[5] Europeansmartcities 4.0. Technische Universität Wien. http://www.smart-cities.eu/?cid=01&ver=4

[6] Рейтинг устойчивого развития городов Российской Федерации. Sustainable growth management agency” (ООО «Агентство ЭС ДЖИ ЭМ»). http://agencysgm.com/projects/Рейтинг%20устойчивого%20 развития-2015.pdf

определяют лидерские позиции города в рейтинге устойчивого развития. Главным критерием данного фактора выступает сбалансированность параметрических показателей. Разбалансировка отрицательно влияет на все стороны устойчивого развития. Ежегодно в названном рейтинге лидирующие позиции занимают Москва, Санкт-Петербург и Уфа. Эти муниципальные образования показывают высокие значения в разрезе социально-экономического развития городов Российской Федерации. В качестве преимущества данного рейтинга можно выделить параллельный анализ показателей социально-экономического развития муниципального образования, а в качестве недостатков можно отметить масштабность, усложненность и непонятность параметрических показателей. Представленные в нем рейтинговые показатели формальны, носят обособленный характер по отношению к технологиям Smart City. Так, например, отсутствие понимания включения в рейтинг показателя «доступность дошкольного образования» или наличие ряда показателей, оценивающих одно направление (в блоке демография и население – коэффициенты естественного прироста, демографического прироста, естественной убыли населения).

Третьей выступает система показателей умных городов, разработанная Европейской экономической комиссией Организации Объединенных Наций [7]. Степень готовности городов к внедрению технологий Smart City в данной системе показателей оценивается через определение инновационности города, использование информационно-коммуникационных технологий и других средств для повышения качества уровня жизни. Кроме того, данная система показателей учитывает эффективность деятельности и услуг, оказываемых в городе, конкурентоспособность при обеспечении удовлетворения потребностей настоящего и будущих поколений в экономических, социальных, культурных и природоохранных аспектах. Представленная система включает три блока со следующими показателями:

Блок 1. Экономика: инфраструктура информационно-коммуникационных технологий; инновационная активность; занятость; электронная торговля и отношение экспорт/импорт; производительность; городская инфраструктура (водоснабжение, электроснабжение, транспорт, эксплуатация зданий и т. д.).

Блок 2. Окружающая среда: качество воздуха; шум; водоснабжение; энергетика; биоразнообразие; качество окружающей среды.

Блок 3. Общество: образование; здравоохранение; культурная сфера; социальная вовлеченность; обеспеченность жилым фондом.

Преимуществом данной системы является подробное описание оценки расчетов представленных показателей. В качестве недостатка можно указать лишь применимость к системе статистической оценки европейских стран.

Проанализировав некоторые имеющиеся на сегодняшний день системы и рейтинги оценки готовности муниципальных образований к внедрению технологий Smart City, попробуем предложить авторский подход к исследованию данного вопроса. Авторское мнение по формированию оценки готовности муниципальных образований к внедрению технологий Smart City заключается в простоте и доступности применения данной модели. В соответствии с этим считаем необходимым сформировать авторскую модель оценки готовности муниципальных образований к внедрению технологий Smart City, использовав элементы имитационного моделирования.

4 Модель оценки готовности муниципальных образований к внедрению технологий Smart City

Проведенное исследование базируется на общенаучных методах анализа, которые включают формально-логический и аналитический способы исследования. Методы, используемые для решения поставленной проблемы в рамках сформулированных задач исследования, определяются закономерностями эмпирического развития данной проблематики и включают: метод моделирования, графический метод, статистический метод. Исходя из этого, используемые в исследовании методы позволяют осуществить процесс имитационного моделирования данных для оценки готовности муниципальных образований к внедрению технологий Smart City.

Авторская модель оценки готовности муниципальных образований к внедрению технологий Smart City сформирована на основе имитационной программы AnyLogic (адаптирована под графические изображения Microsoft Word) и представлена на рис. 1.

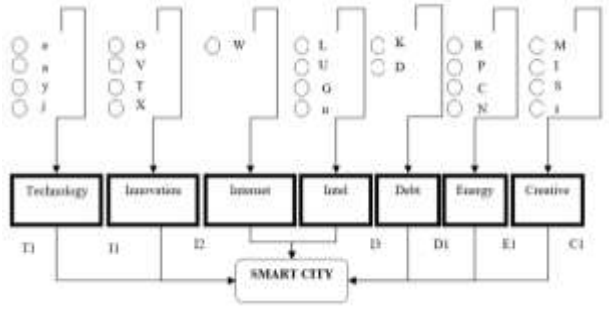


Рисунок 1. Модель оценки готовности городов к внедрению технологий Smart City

Программа имитационного моделирования AnyLogic позволяет разработать модели на основе

[7] Показатели «умных» устойчивых городов, разработанные ЕЭК ООН–МСЭ. Европейская экономическая комиссия Организация Объединенных Наций. http://www.unece.org/fileadmin/DAM/hlm/documents/2015/ECE_HBP_2015_4.ru.pdf

различных методов визуализации данных. Например, дискретно-событийном и агентном методе исследования. Модель оценки готовности городов к внедрению технологий Smart City сформирована на основе семи ключевых показателей (накопителей), включающих в себя параметры распределения, которые позволяют определить итоговый показатель готовности муниципального образования к внедрению технологий Smart City. Выбор показателей, используемых в модели, обусловлен необходимостью оценки уровня инфраструктурно-технологического развития города как ключевого направления внедрения технологий Smart City. В соответствии с этим автором предложены следующие параметрические показатели.

Показатель технологичности производства в муниципальном образовании (Technology)

$$T_p = \frac{e + y + j}{a},$$

где a – общее количество предприятий в муниципальном образовании, e – количество предприятий, проводивших модернизацию не позднее 2007 года, y – количество предприятий, проводивших модернизацию не позднее 2012 года, j – количество предприятий, проводивших модернизацию не позднее 2015 года.

Показатель инновационности городской инфраструктуры (Innovations)

$$I_i = \frac{O}{V} + \frac{T}{X},$$

где O – объем работ, выполненный по замене объектов инновационной инфраструктуры, V – объем работ, требующейся для замены всей инфраструктуры на территории муниципального образования, T – объем инновационной продукции, произведенный в инкубаторах, технопарках и иных инновационных предприятиях муниципального образования, X – объем продукции, произведенный на всех предприятиях муниципального образования.

Показатель интернетизации муниципального образования (Internet)

$$In = \frac{W}{100\%},$$

где W – показатель полного покрытия территории интернетом.

Показатель интеллектуализации городской среды (Intel)

$$I_r = \frac{L + V + G}{n},$$

где L – количество созданных инновационных продуктов, V – количество, зарегистрированных патентов, G – количество выигранных грантов, конкурсов, олимпиад.

Показатель финансовой независимости городского бюджета (Debt)

$$F_n = \frac{K}{D},$$

где K – муниципальный долг, D – доходы бюджета муниципального образования.

Показатель энергоэффективности городской среды (Energy)

$$E_f = \frac{R}{P} + \frac{C}{N},$$

где R – потребление топливно-энергетических ресурсов предприятиями муниципального образования, P – произведенная и отгруженная продукция (товары, работы, услуги) с использованием энергоресурсов, C – стоимость потребляемых энергоресурсов населением, N – население муниципального образования.

Показатель внедрения креативных технологий в функциональное пространство города (Creative)

$$K_t = \frac{M + I + S}{s},$$

где M – количество созданных медиа ресурсов на территории муниципального образования за последние три года, I – реализация проектов индустрии развлечения на территории муниципального образования за последние три года, S – зарегистрированные объекты социального предпринимательства на территории муниципального образования за последние три года, s – зарегистрированные субъекты бизнеса на территории муниципального образования за последние три года.

Все итоговые показатели являются потенциалами. Конечным результатом оценки готовности российских городов к внедрению технологий Smart City является группировка городов на основе следующих критериев:

- готовы к внедрению технологий Smart City (критериальный диапазон $3{,}7(-0{,}2)\langle n)$, n – значение итогового показателя группировки городов по степени готовности к внедрению технологий Smart City;
- средняя готовность к внедрению технологий Smart City (критериальный диапазон $3{,}7(-0{,}2)\langle n\langle 2{,}5(-0{,}3)\rangle$, n – значение итогового показателя группировки городов по степени готовности к внедрению технологий Smart City;
- удовлетворительная готовность к внедрению технологий Smart City (критериальный диапазон $2{,}5(-0{,}4)\langle n\langle 1{,}95(-0{,}4)\rangle, n$ – значение итогового показателя группировки городов по степени готовности к внедрению технологий Smart City;
- не готовы к внедрению технологий Smart City (критериальный диапазон $1{,}95(-0{,}5)\rangle n$), n – значение итогового показателя группировки городов по степени готовности к внедрению технологий Smart City.

Представленные выше диапазоны сформированы в соответствии с наивысшими и наименьшими

значениями каждого показателя, участвующего в определении конечного значения. Максимальный и минимальный уровни показателей, используемых в исследовании, представлены в Таблице 2.

Таблица 2 Наивысшие и наименьшие значения показателей, используемых для определения готовности муниципальных образований к внедрению технологий Smart City[8]

Показатели	Готов к внедрению		Средняя готовность		Удовлетворительная готовность		Не готовы
	max	min	max	min	max	min	max
1 (+)	1	0,6	0,6	0,3	0,2	0,2	0,2
2 (+)	1	0,6	0,4	0,4	0,4	0,3	0,3
3 (+)	1	0,6	0,5	0,3	0,3	0,3	0,3
4 (+)	1	0,7	0,5	0,3	0,3	0,3	0,3
5 (-)	1	0,2	0,8	0,3	0,4	0,2	0,2
6 (+)	1	0,6	0,5	0,4	0,4	0,3	0,3
7 (+)	1	0,6	0,6	0,4	0,4	0,35	0,35
Итог	6	3,7	3,7	2,5	2,5	1,95	1,95

Отметим, что данные диапазоны можно определить не только на основе пробных вариаций, но и при помощи программы PyQt. PyQt – набор «привязок» графического фреймворка Qt для языка программирования Python, выполненный в виде расширения Python.

В рамках имитационной модели оценки готовности городов к внедрению технологий Smart City важным компонентом выступает итоговый индикатор – муниципальное образование. На Рис. 1 данный индикатор (ключевой накопитель) обозначен как SmartCity. В аспекте оценки готовности городов к внедрению технологий Smart City ключевой накопитель приобретает название в соответствии с исследуемым объектов. Для более наглядного демонстрирования данного тезиса применим модель оценки готовности муниципальных образований к внедрению технологий Smart City на примерах Москвы – города федерального значения и Орла – муниципального образования. Проведя оценку готовности, мы установили, что муниципальное образование город Орел относится к критериальной группе «не готовы к внедрению технологий Smart City» (значение 1,837), а город федерального значения Москва входит в группу «готовы к внедрению технологий Smart City» – значение 5,473 (рисунки 2 и 3). Представленная модель позволяет определить не только готовность муниципального образования к внедрению технологий Smart City, но и выявить сегменты (треки), замедляющие процесс перехода в более высшую по уровню критериальную группу. В соответствии с тем, что город федерального значения Москва входит в группу «готовы к внедрению технологий Smart City», замедляющие процессы перехода в более высшую по уровню критериальную группу определим только для муниципального образования города Орел.

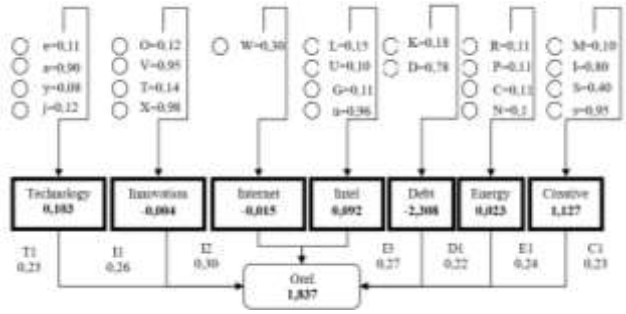


Рисунок 2 Модель готовности к внедрению технологий Smart City муниципального города Орел

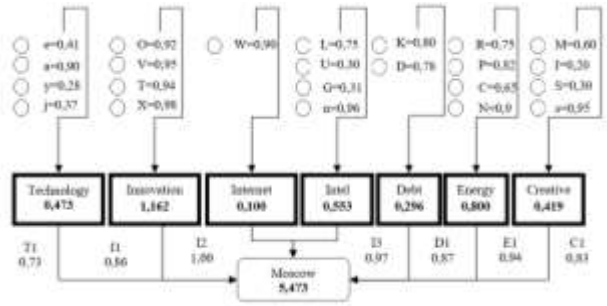


Рисунок 3. Модель готовности к вндрению технологий Smart City города федерального значения Москва

Так, муниципальное образование город Орел находится в критериальной группе «не готовы к внедрению технологий Smart City» в связи с высокой финансовой зависимостью городского бюджета (Debt), отсутствием инновационного развития городской инфраструктуры (Innovations), неполным покрытием территории интернетом (Internet). Для решения представленных выше проблем требуется реализация эффективных управленческих мероприятий по данным направлениям, которые позволят городу Орел приблизиться к группе муниципальных образований, «готовых к внедрению технологий Smart City».

Таким образом, представленная модель оценки готовности муниципальных образований к внедрению технологий Smart City позволит: во-первых, оперативно определить уровень развития муниципального образования, готового к внедрению Smart-технологий; во-вторых, выявить основные проблемы и барьеры, стоящие перед муниципальными образованиями, входящими в критериальную группу «не готовы к внедрению технологий Smart City»; в-третьих, применить соответствующую модель городского развития для реализации Smart-проектов, позволяющих улучшить социально-экономические показатели муниципального образования.

[8] Значения показателей: технологичности производства в муниципальном образовании (1), инновационности городской инфраструктуры (2), интернетизации муниципального образования (3), интеллектуализации городской среды (4), финансовой независимости городского бюджета (5), энергоэффективности городской среды (6), внедрения креативных технологий в функциональное пространство города (7).

Литература

[1] Akaslan, D., Taşkln, S.: An Analogy Between Womb and Home for Supporting the Aspects of Smart Cities. 4th Int. Istanbul Smart Grid Congress and Fair, IEEE Press, New York (2016). doi: 10.1109/SGCF.2016.7492438

[2] Barriga, J.K.D., Romero, C.D.G., Molano, J.I.R.: Proposal of a Standard Architecture of IOT for Smart Cities. Communications in Computer and Information Science, pp. 77-89 (2016). doi: 10.1007/978-3-319-42147-6_7

[3] De Domenico, M., Arenas, A., Lima, A., González, M.C.: Personalized Routing for Multitudes in Smart Cities. EPJ Data Science. 1, pp. 1-11 (2015). doi: 10.1140/epjds/s13688-015-0038-0

[4] Glebova, I.S., Yasnitskaya, Y.S., Maklakova, N.V.: Possibilities of "Smart City" Concept Implementing: Russia's Cities Practice. Mediterranean J. of Social Sciences, 12, pp. 129-133 (2014). doi: 10.5901/mjss.2014.v5n12p129

[5] Ishkineeva, G., Ishkineeva, F., Akhmetova, S.: Major Approaches Towards Understanding Smart Cities Concept. Asian Social Science, 5, pp. 70-73 (2015). doi: 10.5539/ass.v11n5p70

[6] Khatoun, R., Zeadally, S.: Smart Cities: Concepts, Architectures, Research Opportunities. Association for Computing Machinery. Communications of the ACM, 8, pp. 46-57 (2016). doi: 10.1007/978-3-319-23440-3_7

[7] Khorov, E., Gushchin, A., Safonov, A.: Distortion Avoidance While Streaming Public Safety Video in Smart Cities. Lecture Notes in Computer Science, 9305, pp. 89-100 (2015). doi: 10.1007/978-3-319-23440-3_7

[8] Medvedev, A., Fedchenkov, P., Zaslavsky, A., Anagnostopoulos, T., Khoruzhnikov, S.: Waste Management as an IOT-Enabled Service in Smart Cities. Lecture Notes in Computer Science, 9247, pp. 104-115 (2015). doi: 10.1007/978-3-319-23126-6_10

[9] Merlino, G., Bruneo, D., Longo, F., Puliafito, A., Distefano, S.: Software Defined Cities: a Novel Paradigm for Smart Cities through IOT Clouds. 12th IEEE Int. Conf. on Ubiquitous Intelligence and Computing, pp. 909-916. IEEE Press, New York (2015). doi: 10.1109/UIC-ATC-ScalCom-CBDCom-IoP.2015.174

[10] Poxrucker, A., Bahle, G., Lukowicz, P.: Simulating Adaptive, Personalized, Multi-modal Mobility in Smart Cities. Lecture Notes of the Institute for Computer Sciences, Social-Informatics and Telecommunications Engineering, 166, pp. 113-124 (2016). doi: 10.1007/978-3-319-33681-7_10

[11] Zhuhadar, L., Thrasher, E., Marklin, S., de Pablos, P.O.: The Next Wave of Innovation – Review of Smart Cities Intelligent Operation Systems. Computers in Human Behavior, 66, pp. 273-281 (2017). doi: 10.1016/j.chb.2016.09.030

Модифицированный коэффициент корреляции

© Т.О. Дюкина

Санкт-Петербургский государственный университет,
Санкт-Петербург, Россия

dtospb@mail.ru t.dukina@spbu.ru

Аннотация. Статья посвящена рассмотрению показателя – коэффициента корреляции Пирсона, его положительным и отрицательным сторонам применения для анализа динамики и связи между явлениями, а также последующей его модификации. Модификация коэффициента корреляции осуществлена на основе замены способа расчета элементов формулы: средних значений. Осуществлена апробация предложенного модифицированного коэффициента корреляции и доказано его преимущество в более точной оценке тесноты связи между вариацией исследуемого фактора и изменением показателя, характеризующего стабильность налоговой системы страны, на эмпирических данных.

Ключевые слова: коэффициент корреляции Пирсона, модифицированный коэффициент корреляции, средняя арифметическая, средняя геометрическая, вариация, динамика, оценка тесноты связи.

The Modified Correlation Coefficient

© Tatiana Dyukina

St. Petersburg State University,
St. Petersburg, Russia

dtospb@mail.ru t.dukina@spbu.ru

Abstract. Article is devoted to consideration of an indicator – coefficient of correlation of Pearson, to his positive and negative sides of application for the analysis of dynamics and communication between the phenomena, and also the subsequent its modification. Modification of the correlation coefficient is carried out on the basis of replacement of a way of calculation of elements of a formula to be performed on the average values. Approbation of the modified correlation coefficient has been carried out and its merits revealed in more exact assessment of the closeness of links between variation of a studied factor and change of the indicator characterizing stability of the country tax system on empirical data have been shown.

Keywords: Pearson correlation coefficient, modified correlation coefficient, arithmetic average, geometric average, variation, dynamics, assessment of closeness of links.

1 Введение

Сегодня вопросам состояния, развития, а также совершенствования статистических методов начинает уделяться повышенное внимание. Это не случайно, так как именно статистические методы предоставляют широкие возможности своевременного и полного анализа разнообразных данных и получения в результате их обработки качественных выводов.

Исследования, которые посвящены решению не только методологических вопросов статистического анализа в различных сферах экономики и общества, но и оценке универсальности и специализации методов, систематизации опыта применения статистических методов при решении различного рода практических задач, а также развитию и созданию новых методов анализа данных, встречались чаще в прошлом столетии по сравнению с сегодняшним днем. В настоящее время такие исследования являются относительно большой редкостью. Таким образом, вопросы совершенствования статистических методов, в том числе отдельных статистических показателей, приобретают еще большую актуальность в свете обозначенных аспектов.

Данная статья посвящена совершенствованию методики расчета одного из наиболее употребляемых в статистической практике анализа данных различных показателей – для оценки тенденции ряда динамики и тесноты связи между показателями. Для статистического анализа тенденции ряда динамики, а также тесноты связи между вариацией исследуемого фактора и изменением изучаемого показателя применение находит широко известный показатель: коэффициент корреляции Пирсона.

Труды XIX Международной конференции «Аналитика и управление данными в областях с интенсивным использованием данных» (DAMDID/ RCDL'2017), Москва, Россия, 10–13 октября 2017 года

2 Анализ степени исследования проблемы

2.1 Показатели, применяемые для измерения стабильности (устойчивости) тенденции ряда динамики

Действительно, для измерения стабильности (устойчивости) тенденции ряда динамики среди рекомендованных к применению показателей: коэффициента корреляции рангов Ч. Спирмена (C.E. Spearman) [9, с. 345; (Spearman)] и соотношения между среднегодовым абсолютным изменением и средним квадратическим (либо линейным) отклонением уровней от тренда [9, с. 347] индексу корреляции, показывающему степень сопряженности колебаний фактических уровней с колебаниями теоретических уровней, происходящих под влиянием комплекса основополагающих факторов, и представляющему собой коэффициент корреляции Пирсона (Pearson), или иначе, линейный коэффициент корреляции [7, с. 475] отводится одно из самых важных, можно сказать, эпохальных мест.

Здесь следует акцентировать внимание на том, что, во-первых, коэффициент корреляции Пирсона рекомендуется использовать исключительно в случаях линейной связи. В случаях нелинейной связи, которые встречаются наиболее часто, применение данного показателя нежелательно. Во-вторых, слабым местом данного показателя является его неверное реагирование на выбросы: результаты измерения, выделяющиеся из общей совокупности (слишком большие или малые значения) могут способствовать большим значениям данного показателя. В таком случае, они означают высокую степень сопряженности колебаний фактических уровней с колебаниями теоретических уровней, происходящих под влиянием комплекса основополагающих факторов. В-третьих, в случаях, когда одна из двух переменных не является нормально распределенной (а, как показывает анализ множества эмпирических данных, имеющих экономическую природу, большинство таких данных собственно и не являются нормально распределенными), а также в случаях, когда одна из двух переменных имеет порядковую шкалу измерения, коэффициент корреляции Пирсона неприменим. В этих случаях рекомендуется использовать только ранговые коэффициенты Спирмена и Кендалла.

2.2 Показатели, используемые для количественной оценки влияния отдельных факторов на анализируемый показатель

Для определения количественной оценки влияния отдельных факторов на анализируемый показатель в настоящее время имеется возможность применять различные методы: индексный анализ, дисперсионный анализ, корреляционно-регрессионный, эконометрический анализ и другие. В последнее время наибольшее распространение в научных исследованиях получили методы корреляционно-регрессионного и эконометрического анализа. Обозначенные методы для определения количественной оценки влияния отдельных факторов на уровень стабильности налоговой системы страны требуют осмотрительного, пунктуального и вдумчивого применения, поскольку в процессе их применения могут возникать целый ряд еще неразрешенных в науке в полном объеме проблем:

- использование неполного комплекта влияющих факторов;
- построение моделей, которые содержат ненаблюдаемые факторы;
- ложная причинно-следственная связь, в том числе возникающая из-за употребления в анализе замещающих факторов [1].

Следовательно, широкое применение рассматриваемого в настоящей статье показателя – коэффициента корреляции Пирсона, особенно без учета его особенностей и специфики применения, может привести к неверным расчетам и выводам. Отмеченное становится особенно актуальным в случаях нелинейности развития анализируемых показателей, характеризующих экономическую среду.

3 Методологические вопросы разработки модифицированного коэффициента корреляции

3.1 Особенности экономической среды

Многие экономисты, в числе которых Дж. Кейнс, считают экономическую среду непредсказуемой и изменчивой [5]: «экономическая среда на протяжении некоторого периода времени должна оставаться неизменной и однородной во всех значимых отношениях, за исключением колебаний тех факторов, которые рассматриваются отдельно» [3]. «Но быть уверенными, что такие условия сохранятся в будущем, даже если они обнаруживаются в прошлом, нельзя», – заключает ученый [3].

Действительно, большинство экономических переменных (факторов) взаимодействуют посредством многообразных нелинейных зависимостей. Однако арсенал эконометрической науки сегодня довольно богат, что позволяет успешно решать проблемные вопросы при моделировании социально-экономических процессов и явлений.

3.2 Коэффициент корреляции Пирсона

Как уже отмечено выше, в случае линейных зависимостей широкое применение для определения тесноты связи находит коэффициент корреляции Пирсона (см. формулу 1).

$$K_{xy} = \sqrt{\frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{n}(x_i - \bar{x})^2 \sum_{i=1}^{n}(y_i - \bar{y})^2}} \quad (1)$$

где K_{xy} – значение коэффициента корреляции Пирсона, $\bar{x}$, $\bar{y}$ – средние значения уровней показателя, рассчитываемые по формуле арифметической средней [8, с. 224].

3.3 Модифицированный коэффициент корреляции

Следует учитывать, что довольно большой объем факторов, вариация которых оказывает влияние на изменение анализируемого показателя, подчиняется законам распределения, характер которых отличен от нормального распределения. Представляется, что среднее значение показателя, рассчитанное по формуле арифметической средней, в этих распределениях не является истинным. В этом случае расчет среднего значения исследуемого показателя по геометрической средней, учитывающим большой разброс значений показателя, представляется более корректным. Вследствие этого, полагаем возможным осуществить модификацию коэффициента корреляции Пирсона посредством введения в нее вместо среднего значения, определяемого по формуле арифметической средней, среднего значения, рассчитанного по формуле геометрической средней Пирсона (см. формулу 2).

$$K_{xy}{}^{M} = \sqrt{\frac{\sum_{i=1}^{n}(x_i - \bar{x}_{геом})(y_i - \bar{y}_{геом})}{\sum_{i=1}^{n}(x_i - \bar{x}_{геом})^2 \sum_{i=1}^{n}(y_i - \bar{y}_{геом})^2}} \quad (2)$$

где $K_{xy}{}^{M}$ – значение модифицированного коэффициента корреляции, $\bar{x}_{геом}$, $\bar{y}_{геом}$ – средние значения уровней фактора и результативного показателя, определяемые по формуле геометрической средней.

При исследовании совокупностей с качественно разнородными признаками на первый план выступает именно нетипичность средних показателей. Средняя геометрическая величина позволяет осуществить обобщение качественно разнородных значений признаков системных пространственных совокупностей или статистических совокупностей, представленных в динамике (во времени). Она, обнаруживая общие свойства исследуемых совокупностей, которые присущи всем единицам соответствующих совокупностей, позволяет выявить общие закономерности, обусловленные общими причинами, а также избежать случайных влияний.

При модификации коэффициента корреляции Пирсона была выбрана именно средняя геометрическая величина, так как она позволяет наилучшим образом осуществить обобщение значений признака в исследуемой совокупности не только в случаях наличия экстремальных значений отдельных единиц изучаемой статистической совокупности, но и в случаях распределений, принимающих характер, отличающийся от нормального закона распределения.

На наш взгляд, замена средней арифметической величины при модификации коэффициента корреляции Пирсона на другие статистические величины (например, медианное значение, модальное значение, а также иные робастные величины) не рациональна, поскольку не позволит в должной мере обеспечить устойчивость меры среднего. Кроме того, стоит отметить тот факт, что использование Пирсоном модального и медианного значений в известных формулах асимметрии распределений не сделало их более совершенными, наоборот, они общепризнанно считаются весьма приблизительными и довольно часто показывают некорректные значения этого показателя.

Возможности модифицированного коэффициента корреляции (по сравнению с коэффициентом корреляции Пирсона) более обширны: его можно применять для оценки тесноты связи между вариацией исследуемого фактора и изменением показателя, характеризующего анализируемый показатель, в случаях, когда характер распределений исследуемого фактора и (или) показателя, его характеризующего, отличается от закона нормального распределения (поскольку применение среднего значения, рассчитанного по геометрической средней, позволяет корректно учитывать большой разброс значений показателя в распределениях, отличных от нормального закона распределения). В результате модифицированный коэффициент корреляции позволит наиболее точно определять силу влияния вариации фактора на изменение исследуемых показателей.

4 Апробация модифицированного коэффициента корреляции

4.1 Данные и выборка

В настоящем исследовании осуществлена также апробация предложенного модифицированного коэффициента корреляции и эмпирически доказано его преимущество, заключающееся в более точной оценке тесноты связи между вариацией исследуемого фактора и изменением показателя, характеризующего анализируемый показатель.

В качестве исследуемого показателя был выбран показатель, характеризующий стабильность налоговой системы нашей страны – средняя фактическая налоговая нагрузка на одного налогоплательщика по налогам, сборам и иных обязательным платежам в бюджетную систему Российской Федерации, а в качестве фактора – уровень заболеваемости всего населения с диагнозом, установленным впервые в жизни на 1000 человек населения.

Исследования осуществлены в динамике за период 2010-2014 гг. на основе официальных статистических данных в разрезе субъектов Российской Федерации, формируемых Федеральной налоговой службой России и Федеральной службой государственной статистики.

Средняя фактическая налоговая нагрузка на одного налогоплательщика по налогам, сборам и иных обязательным платежам в бюджетную систему Российской Федерации определена на основе данных ФНС [6]. Федеральная налоговая службой России представляет данные в свободном доступе в целом по Российской Федерации и в разрезе ее субъектов за период с 2007 г. по настоящее время в формах статистической налоговой отчетности.

Уровень заболеваемости всего населения с диагнозом, установленным впервые в жизни на 1000 человек населения, рассчитан на основе данных Федеральной службы государственной статистики [2, 4].

Поскольку данные по исследуемым показателям были взяты в разрезе субъектов Российской Федерации, следовательно, в работе был применен сплошной метод исследования.

4.2 Эмпирические результаты исследования

Предварительно был осуществлен анализ исследуемого показателя, характеризующего стабильность налоговой системы нашей страны – средней фактической налоговой нагрузки на одного налогоплательщика по налогам, сборам и иным обязательным платежам в бюджетную систему Российской Федерации на основе расчета показателей центра, структуры, степени вариации и типа распределения и установлен характер распределения субъектов РФ налоговой системы по показателю средней фактической налоговой нагрузки на одного налогоплательщика по налогам, сборам и иным обязательным платежам в бюджетную систему Российской Федерации за период 2010-2014 гг. (см. Таблицу 1).

Таблица 1 Показатели центра, структуры, степени вариации и типа распределения средней фактической налоговой нагрузки на одного налогоплательщика по налогам, сборам и иным обязательным платежам в бюджетную систему Российской Федерации за период 2010-2014 гг.

Показатели	Годы				
	2010	2011	2012	2013	2014
1	2	3	4	5	6
Средняя арифметическая	62	81	91	91	104
Средняя геометрическая	31	37	42	43	49
Медианное значение	27	31	36	36	41
Размах вариации	901	1 179	1 340	1 224	1 311
Среднее линейное отклонение	61	83	92	91	104
Среднее квадратическое отклонение	141	191	222	217	238
Коэффициент вариации, %	225,7	237,0	244,1	237,9	229,2
Коэффициент асимметрии	4,79	4,72	4,77	4,71	4,57
Коэффициент эксцесса	23,34	22,35	22,37	21,45	20,36

Источник: рассчитано автором

Анализ средних и медианных значений изучаемого показателя за период 2010-2014 гг. (рассчитанных по несгруппированным данным), свидетельствует об их стабильном увеличении на протяжении всего исследуемого периода, что означает положительные изменения исследуемого показателя на макроуровне и, как следствие, направленность изменений в сторону стабильного развития налоговой системы в Российской Федерации (следует отметить, что здесь сказывается влияние инфляционного фактора). Однако все показатели вариации, а также коэффициенты асимметрии и эксцесса (рассчитанные по несгруппированным данным), являются более тонким инструментом, позволяющим учитывать влияние случайных факторов на исследуемый показатель, и указывают на постоянную и довольно существенную вариацию значений рассматриваемого показателя. Анализ коэффициента вариации в исследуемом периоде показал также, что в РФ совокупности субъектов по исследуемому показателю за период 2010-2014 гг., чрезвычайно неоднородные, вариация по субъектам РФ значительная, так как превышает не только 33%, но и 100%, что свидетельствует о крайней нестабильности налоговой системы в пространственном аспекте за анализируемый период. Следует отметить уменьшение значений характеристик распределения (коэффициентов асимметрии и эксцесса) по исследуемому показателю в 2014 г. по сравнению с 2010 г., пусть и незначительное, но, тем не менее, это указывает на позитивные изменения, происходящие в развитии налоговой системы страны.

Среднее значение анализируемого показателя, рассчитанного по формуле геометрической средней в два и более раз меньше, чем аналогичное значение, рассчитанное по формуле арифметической средней,

на протяжении всего исследуемого периода. При этом именно значения показателя, рассчитанные по формуле геометрической средней, наиболее приближены к медианным значениям, что косвенно подтверждает их преимущество в выявлении истинного среднего значения в исследуемой совокупности.

Таким образом, анализ показателей центра, структуры, степени вариации и типа распределения исследуемого показателя за период 2010-2014 гг. позволяет констатировать, что распределение изучаемого показателя на протяжении всего рассматриваемого периода имеет характер гиперэкспоненциального распределения.

На основе данных показателя, характеризующего стабильность налоговой системы нашей страны – средней фактической налоговой нагрузки на одного налогоплательщика по налогам, сборам и иных обязательным платежам в бюджетную систему Российской Федерации, и его фактора – уровня заболеваемости всего населения с диагнозом, установленным впервые в жизни на 1000 человек населения – в динамике за каждый год периода с 2010 по 2014 гг. был рассчитан модифицированный коэффициент корреляции, а также модифицированный коэффициент детерминации (рассчитываемый возведением в квадрат модифицированного коэффициента корреляции). Результаты расчетов эмпирических исследований по расчету коэффициентов корреляции и детерминации (в том числе модифицированных) представлены в таблице 2.

Таблица 2 Коэффициенты корреляции и детерминации взаимосвязи средней фактической налоговой нагрузки на одного налогоплательщика по налогам, сборам и иных обязательным платежам в бюджетную систему Российской Федерации и уровня заболеваемости всего населения с диагнозом, установленным впервые в жизни на 1000 человек населения за период 2010-2014 гг.

Показатели	Годы				
	2010	2011	2012	2013	2014
1	2	3	4	5	6
Коэффициент корреляции	0,628	0,600	0,572	0,524	0,457
Коэффициент детерминации	0,395	0,360	0,327	0,274	0,209
Модифицированный коэффициент корреляции	0,410	0,417	0,323	0,203	0,184
Модифицированный коэффициент детерминации	0,168	0,174	0,104	0,041	0,033

Источник: рассчитано автором

Для более наглядного представления информации представим полученные коэффициенты корреляции на графике (см. Рисунок 1) .

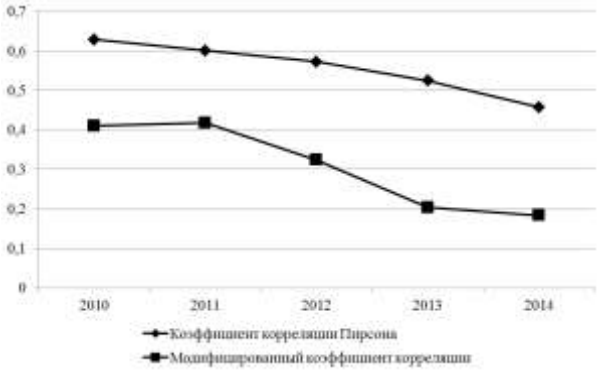


Рисунок 1 Коэффициенты корреляции взаимосвязи средней фактической налоговой нагрузки на одного налогоплательщика по налогам, сборам и иным обязательным платежам в бюджетную систему Российской Федерации и уровня заболеваемости всего населения с диагнозом, установленным впервые в жизни на 1000 человек населения за период 2010-2014 гг.

5 Заключение

На основе анализа данных таблицы 2 и рис.1 можно констатировать, что за весь рассматриваемый период уровень значений модифицированного коэффициента корреляции по сравнению с коэффициентом корреляции Пирсона является существенно более низким. Это означает существенно более низкую (в данном случае, в отдельные годы даже более чем в два раза) взаимосвязь вариации средней фактической налоговой нагрузки на одного налогоплательщика по налогам, сборам и иным обязательным платежам в бюджетную систему Российской Федерации и уровня заболеваемости всего населения с диагнозом, установленным впервые в жизни на 1000 человек населения за период 2010-2014 гг.

Аналогичный вывод можно сделать и по рассчитанным коэффициентам детерминации: за весь анализируемый период уровень значений модифицированного коэффициента детерминации оказался меньше, чем у коэффициента корреляции Пирсона.

Кроме того, выявленный характер взаимосвязи вариации исследуемых показателей заметно отличается, что особенно хорошо заметно на графике. Модифицированный коэффициент корреляции имеет более высокие темпы снижения для исследуемых эмпирических показателей по сравнению с коэффициентом корреляции Пирсона, что является следствием более точного учета влияния дисбаланса анализируемой экономической системы, в которой были выявлены нелинейные процессы развития.

Следовательно, использование модифицированных коэффициентов корреляции и детерминации позволяет получить, по нашему мнению, более точную оценку взаимосвязи изменений средней фактической налоговой нагрузки на одного налогоплательщика по налогам, сборам и иных обязательным платежам в бюджетную систему Российской Федерации и уровня заболеваемости всего населения с диагнозом, установленным впервые в жизни на 1000 человек населения.

Литература

[1] Hendri D. Econometrics: alchemy or science? Ekovest, No. 2. pp. 172 – 196 (2003)

[2] Incidence of the population on the main classes of diseases. http://www.gks.ru/wps/wcm/connect/rosstat_main/rosstat/ru/statistics/population/healthcare/#

[3] KeynesJ. M. Method of professor Tinbergen. Economy questions. No. 4. P. 28 (2007)

[4] Population. http://www.gks.ru/wps/wcm/connect/rosstat_main/rosstat/ru/statistics/population/demography/#

[5] Rozmainsky I. Methodological bases of the theory of Keynes and his "dispute on a method" with Tinbergen. Economy questions. No. 4. pp. 25-36 (2007)

[6] Summary reports in general on the Russian Federation and in a section of subjects of the Russian Federation. https://www.nalog.ru/rn78/related_activities/statistics_and_analytics/forms/

[7] The tendency of property stratification only accrues. Experts warn about danger of social explosion in Russia because of property stratification [An electronic resource]. http://www.newizv.ru/economics/2014-10-17/209143-tendencija-imushestvennogo-rassloenija-tolko-narastaet.html. – Zagl. from the screen (2014)

[8] Theory of statistics. Under the editorship of the prof. G. L. Gromyko. 2nd prod., reslave. and additional Moscow. 476 p. (2006)

[9] Yeliseyeva I. I., Yuzbashev M. M. General theory of statistics. 4 prod., reslave and additiona. Moscow. 480 p. (1999)

Towards Framework for Discovery of Export Growth Points

© Dmitry Devyatkin[1] © Roman Suvorov[1] © Ilya Tikhomitov [1] © Yulia Otmakhova [2]

[1]Federal Research Center Computer Science and Control of the Russian Academy of Sciences, Moscow, Russia
[2]Novosibirsk State University, Novosibirsk, Russia

devyatkin@isa.ru rsuvorov@isa.ru tih@isa.ru otmakhovajs@yandex.ru

Abstract. Export value of the Russian Federation has been reducing in the latest years, as well as the corresponding relative yield. Most probably, this trend is caused by Russia total export decline together with growth of food export. Thus, it is very important to not only increase export volumes, but also adjust export structure to fit nowadays reality better. The paper presents a computer-aided framework for export growth points discovery. While the full framework is described briefly, more attention is paid to the first sub-task: growth point candidates ranking. The objective of this sub-task is to reveal combinations of commodities and partner countries with high probability of successful export. The method uses open data about international trade flows and production from United Nations databases and modern machine learning methods. The experimental evaluation shows that taking into account retrospective data allows ranking growth point candidates significantly better. Finally, the limitations and the possible directions of future research are discussed.

Keywords: export growth potential, data mining, international trade, customs statistics, open data, machine learning.

1 Introduction

Sanctions pose both difficulties and opportunities for the Russian economy. On the one hand, traditional foreign markets may be restricted or their growth potential may be exhausted. On another hand, exploring new markets may become a fruitful workaround. We believe that modern big data and machine learning technologies should be useful to discover new foreign markets with high probability of growth in the nearest future. We will refer to the pairs of countries and commodities as potential *growth points*. This paper aims on making a step towards finding new growth points using machine learning and open data analysis.

Authors of [1] consider export growth potential as an opportunity to meet the primary demand for a certain product or service. At the same time, the possibility to satisfy the demand arises locally and has a specific territorial, and, therefore, national binding.

There are two possible ways to satisfy growing demands: extensive and intensive. Intensive way implies improving technologies, scientific and engineering solutions and increasing the resource potential and efficiency of management. Therefore, a product may have high export growth potential if it has high added value, robust interbranch relations and stable external demand. In this paper, we propose a framework for discovery of "export growth points". High-level procedure of this framework consists of two main steps: (1) finding candidates for "growth points"; (2) assessing each candidate and discovering difficulties with its implementation. The first step consists in ranking pairs <commodity, foreign market> in such a way so most likely growing pairs appear in the beginning of the list.

In this paper we propose a machine-learning-based method that ranks the "growth point" candidates using features, extracted from historical data from FAOSTAT and UN Comtrade databases [2, 3]. The presented evaluation is preliminary, because it is based on retrospective data. We understand such a weakness and we are going to address it in the future work.

The rest of the paper is organized as follows: in the Section 2 we review the most related works published so far; in Sections 3 and 4 we briefly describe our framework and the task of export growth point candidates ranking; in Section 5 we describe our dataset and present the results of experimental evaluation; in Section 5 we conclude and discuss future work.

2 Related work

Most commonly used approaches to foreign trade modeling include: gravity models, computable general equilibrium models, heuristic ranking models, Markovian models, common statistical approaches (regressions, histograms) for manual analysis of a situation.

The paper [4] presents the empirical evaluation of spatial gravity model of Russian trade. The authors concluded that the spatial variables such as the location of the state border checkpoints have a significant effect on the volume and routes of Russian imports. In [5] authors study factors of export and import value-added trade and suggest some recommendations for management of industrial and trade policy. The techniques proposed in this paper allow to determine main directions of economic policy to expand exports

Proceedings of the XIX International Conference "Data Analytics and Management in Data Intensive Domains" (DAMDID/RCDL'2017), Moscow, Russia, October 10–13, 2017

and improve Russian production structure. Duenas and Fagiolo in their paper [6] concluded that gravity models are poorly suited to predicting the presence of trade relations between some two countries. However such models allow us to accurately estimate and forecast the volume, given the knowledge that such trade relation exists. In [7] researchers use gravity models to investigate the export destinations that could be effectively developed with internal financial support. Experimental work was carried out on the data of food export at the firm-level.

In [8] authors consider Markov models for forecasting the variability of the network of foreign trade financial flows. In [9] an approach for detecting promising areas of export in the sector of both service and goods is proposed. The approach is based on the sequential filtering of potential markets via a number of heuristics, including estimation of the market volume, a level of demand, market openness, etc. In [10] authors studied the relationships between migration flows and foreign trade. They concluded that the trade flows for some products are positively and significantly correlated with migration flows. That feature can be taken into account during analyzing and evaluating the prospects of an export.

In [11] Lall et al. investigated relationship between exports volume and the "complexity" of goods and introduced a metric of "complexity" or "manufacturability" of goods. They mentioned the dependence between the rate of growth of prices on a product and the degree of it manufacturability. This dependence can be used as one of the features for detecting and assessing the export growth potential. Bernard et al. [12] proposed a method for estimating the feasibility of entering the international market for a particular company. They used indicators of the company past activity, including participation in exports, a competitive environment, etc. It is worth noting the weak influence of sectoral state support for exports on the actual volume of exports. In [13] authors considered the relationship of the topology of the international trade network between countries in general with network topologies within each product group. They proposed a methodology for studying the dynamics of changing the structure of several heterogeneous networks that represent trade flows between countries for individual commodity groups. As a result, the most active exporters and importers were detected for separate groups.

In [14] authors try to model the structure and dynamics of the international trade network using the classical methods for solving selecting balls from urns problem. The analysis is carried out at the level of countries and the principle of preferential attachment is implemented ("the rich get richer, the poor get poorer").

In [15] authors propose to model the structure and dynamics of the International Trade Network via the Hamiltonian system. The authors describe the dynamics of the International Trade Network in terms of Hamiltonian, and also make the assumption that the main provisions from the field of statistical physics will also be applicable to modeling the International Trade Network.

Shen et al [16] considered the international trade network at the level of countries and goods. They used flow analysis in graphs and statistics on tops to study the network. The authors draw a number of conclusions related to the specialization of countries, as well as the dominance of developed countries in terms of the diversity of exported products (the principle of preferential accession).

They empirically confirm the fact that food products are mostly traded between the most closely located countries, while high-tech goods are distributed virtually all over the world. Also, the authors detect countries with an anomalous profile of imports, which can talk about a number of economic problems. In [17] authors presented the analysis of export in the service sector on the example of Germany companies. The main goal of the analysis is to determine the dependence of directions and the mode of export on the various features of exported services. They used a non-open dataset from Deutsche Bank. Among other things, the authors detected such heuristics as "exports are more preferable to countries with higher incomes (for countries with lower incomes, an international partnership is more preferable)"; "When selling in more remote countries, international partnership is more profitable."

In [18; 19] researchers developed machine learning models to forecast export dynamics of agricultural products. They compare Support Vector Machines (SVM) and Autoregressive Integrate Moving Average (ARIMA). The experiments showed that SVM achieves significantly smaller error rates.

To sum the review up, we can say that quite extensive efforts have been committed to analyze and predict international trade flows. However, most papers describe fragmentary studies, which are focused on a limited set of factors. Thus, a goal-oriented and comprehensive approach is in high demand.

3 Framework for discovering export growth points

In this section we will try to formalize the problem of export growth points discovery. The objective is to find combinations $<Product_i, Country_j>$, which have the highest unrealized potential for export growth. Also, production and export management of these combinations has to be feasible in the Russian Federation. $Product_i$ is a product or product category to export and $Country_j$ is a country or a group of countries to export to.

We propose to use open data analysis and modern machine learning techniques to find such growth points. The high-level algorithm of our framework consists of the following steps:

1. Construct a list of growth point candidates$<Product_i, Country_j>$. Reorder this list so the candidates with higher likelihood of becoming successful export direction appear earlier.
2. Analyze supply chains which contain commodities from our candidate list. Products

with higher added value should be reviewed first. Consider the product lifecycle (including production, storage, transportation and processing for the selected products) in order to detect the most probable difficulties for each stage of the lifecycle in the context of the Russian Federation. Propose intensive or extensive ways of overcoming them. Products with too many difficulties are removed from the list.

Novelty of our approach consists in maximum possible automation. We can automate step 1 (candidates ranking) and aid step 2. Ranking in Step 1 can be carried out with a predictive machine-learning based model. Step 2 can be highly facilitated by developing a specialized information retrieval system which uses big collections of scientific and engineering documents, such as patents, scientific papers, grant reports. Step 1 is discussed in detail later in this paper. We are going to consider step 2 in future.

4 Data Driven Candidates Ranking

Formally, the problem of candidates ranking is a Learning-To-Rank (LTR) problem. Traditionally, each LTR problem is specified by three components: a set of possible queries, a set of objects and a target metric to optimize. In this work each query is formulated as "Which products to which countries should we try to export to increase budget income, in the context of current macroeconomic situation and our state of industry?". In other words, a query is specified by current economic context (wide or narrow, depends on implementation). Objects that are ranked relative to that query are export growth point candidates or pairs $<Product_i, Country_j>$ (what and where to export).

The main difficulty with LTR problem statement is target metric construction. This metric must reflect the likelihood of success if export of $Product_i$ to $Country_j$ from the Russian Federation will be established. Such a metric cannot be constructed in purely data-driven way, because no database of such cases exists. To overcome this issue, we propose to base on two sources of knowledge: (1) opinion of experts in the field of food market and international trade; (2) retrospective data about dynamics of international trade. On the one hand, retrospective data alone cannot be used to predict future, because the world context is changing and it will almost never become same again. On another hand, experts base on a limited number of factors and limited knowledge (it may be very deep but still limited). Thus, we propose to use experts to take into account factors which are hard to formalize; and retrospective data - to measure prior likelihood of trade flow of $Product_i$ to $Country_j$ to grow.

Taking into account expert opinion requires labeling a training dataset. In this paper we conduct preliminary studies only using retrospective data, due to limitations of time and resources. Experiments with manually annotated datasets will be considered in future.

In other words, in this paper we study only export dynamics prediction. One can dispute that LTR is a reasonable approach to this problem and claim that traditional regression is a better fit. We chose LTR due to three main reasons. The first one is that information about order is more abstract than information about exact increase of trade value or volume (and thus the corresponding predictive model should generalize better). The second reason is that we plan to use LTR in more general case and thus we want to conduct experiments as close to the proposed framework as possible. And the third reason is that we can generate more data to train LTR model and thus try to reduce overfitting.

To facilitate solution of the described LTR problem, we treat it as pairwise ranking problem: we build a regression model, which is given a pair of two export growth point candidates $<Product_1, Country_1>$ and $<Product_2, Country_2>$ returns a difference between export flows for the first and second pair. Generally, such a model operates on a feature set consisting of three major parts: description of global macroeconomic situation; description of trade flows for the first candidate; description of trade flows for the second candidate. Ideally, information about both candidates should also somehow describe prices, competitiveness, quality etc.

The objective of the experimental evaluation in this paper is to verify that retrospective data is useful to compare trade flow dynamics for different commodities and foreign markets. To achieve this goal, we applied ARIMA model as a baseline and also built two machine learning models: "baseline" and "advanced".

4.1 Dataset

We used excerpts from FAOSTAT [2] and UN Comtrade (Comstat) [3] databases from 2011 to 2015 years. The main source of data is Comstat (import, export, re-import, re-export). From FAOSTAT we took information about production volumes. The last year FAOSTAT contains data about is 2014, so 2015 is the last year we could predict for. Full dataset contained 307 million data points.

Due to limited time and computational resources, we conducted experiments only on the 10 most exported from the Russian Federation commodities. Also, we selected 20 countries in the same way. Thus, we got 200 growth points. Surely, in future experiments we should consider much larger set of commodities and countries, not only those well-developed already.

The testbed was set up as follows. All available data were split into two parts: train and test. Train subset contained information about trade from 2013 to 2014. Test subset contained information about only 2015. Each subset consisted of datapoints each representing a pair of export growth point candidates to compare. Features were constructed using "current" and "previous" year. Outcomes were constructed on the base of the "next" year. Thus, in train features were constructed on the base of 2011-2012 (2013 as "next") and 2012-2013 (2014 as "next") and outcomes were constructed on the base of 2013 and 2014 correspondingly. In test subset features

Table 1 Top 5 predicted export growth points and their summary proportion in the total export gain

No	Actual		Predicted					
			ARIMA		Baseline model		Advanced model	
	Partner Country	Commodity	Partner Country	Commodity	Partner Country	Commodity	Partner Country	Commodity
1	Saudi Arabia	Barley	Libya	Barley	Azerbaijan	Potatoes	Italy	Maize
2	China	Soybeans	Spain	Soybeans	Georgia	Maize	Spain	Maize
3	Turkey	Maize	Ukraine	Wheat	Uzbekistan	Wheat	Libya	Maize
4	Azerbaijan	Wheat	Ukraine	Molasses	Ukraine	Potatoes	Spain	Rye
5	Italy	Maize	Kazakhstan	Soybeans	China	Wheat	Ukraine	Molasses
Export gain	$	360059k		11710k		13830k		19197k
	%	**76.2**		**2.4**		**2.9**		**4.0**

were constructed using 2013-2014 and outcomes represented difference in dynamics in 2015. Each subset was symmetric: for each pair <A, B> there was also pair <B, A>. Samples with outcome of 0 were excluded from both subsets.

4.2 Baseline model

The objective of baseline model is to estimate, how accurate candidates can be compared using only knowledge about titles of these candidates. Baseline is implemented as Bernoully Naive Bayes classifier with feature set, consisting only of <$Product_1$, $Country_1$> *(only elements of left hand part of comparison)*. Etalon oucomes for training the baseline model were constructed as $sign(dEV_1 - dEV_2)$, where dEV_i is the first difference of export value of $Product_i$ from the Russian Federation to $Country_i$.

Thus, this classifier estimates prior marginal probability of each candidate to grow faster than each other candidate. This model is very naive and measures skewness of our dataset and most frequent patterns of the Russian Federation international trade.

4.3 «Advanced» model

The objective of this model is to estimate, how much simple context information can improve comparison accuracy. There are several differences from the baseline: the feature set, the machine learning method used and the loss function.

The feature set consists of two parts: historical information about trade of the Russian Federation with $Product_i$ and $Country_i$; and the same information about the second candidate. "Historical information about trade" includes the following basic values from UN Comtrade database: export amount (in tonnes), export value (in USD), export prices (as ratio of value to amount), export monopolization; the same corresponding parameters for re-export, import, re-import. The feature set also contains information about production (from FAOSTAT database). Prior dynamics is taken into account using first order differences and ratios. First order difference (or ratio) is the difference (or ratio) of the value for the current year and that for the previous one. Monopolization (or competitiveness, or concentration) is estimated using Herfindahl index (sum training data was split so that data for each year was used solely either for the train or for evaluation. After best hyperparameters were chosen, the model was refitted using all training data. Finally, we decided to use LightGBM to train that model, because it showed the most promising results. All the results presented for "advanced" model were constructed using LightGBM.

One can notice that we do not explicitly use information about global economic situation. We omitted it from the feature set due to two main reasons: (1) it is very difficult to represent in such a way so a machine learning-based model can take full advantage of it (unclear how to prepare features); (2) some global information is implicitly encoded into difference between production, import and export, and also in monopolization estimates. Surely, explicitly taking into account the global economic situation is very important. We will consider it in next papers.

5 Experimental evaluation

As written before in the paper, the main objective of experimental evaluation is to estimate how much the detailed retrospective data about international trade is useful for the problem of growth point candidate ranking. Because of the nature of the problem, the standard classification or regression scores are not well applicable to measure the prediction quality, i.e. miscomparison of different pairs may have very different significance. Therefore, we used a proportion of the predicted export growth points in the total export gain as the score. In other words, the bigger part of export growth the model detects (the list "%" row in tables), the better the model works. These percent values may be treated as quantitative prediction quality measures.

Table 1 contains the scores for the top 5 actual growth points and for the predicted alternatives. Sum absolute export value growth for the predicted pairs is presented. The last row (%) contains the portion of total growth of export from Russia in 2015, calculated for all growth point candidates (as specified above). From this table one can see that it is nearly impossible to predict short one-year trade flow dynamics without additional information about global economic situation.

A notable difficulty here is high volatility of the product market, while the creation or development of a

food manufacture is a long-term process. Therefore, we think that prediction of averaged, long-term trends would yield a more meaningful ranking.

Advanced model achieved slightly better results than baseline and ARIMA models. From that we conclude that retrospective data is useful to predict flow dynamics. This in turn means that combining open retrospective data about international trade with expert opinions makes much sense in order to maximize both likelihood and novelty.

Table 2 Top 5 predicted commodities and their proportion in the total export gain

No	Actual	ARIMA	Baseline model	Advanced model
1	Barley	Barley	Potatoes	Maize
2	Soybeans	Soybeans	Maize	Rye
3	Maize	Wheat	Wheat	Molasses
4	Wheat	Molasses	Linseed	Soybeans
5	Potatoes	Maize	Rye	Wheat
$	446903k	440272k	137694k	225233k
%	**94.6**	**93.2**	**29.1**	**47.6**

Table 2 presents five commodities with the highest expected growth. The last row (%) contains the portion of total growth. One can see how much Russian food export is non-diversified: 5 commodities occupy more than 90% of total export value growth. Also, we can see that ARIMA predicts commodity dynamic much better than both baseline and advanced model. We think that this is mostly due to inertia of flows: if something grows today, it will most probably grow tomorrow. Again, "advanced" model performed better than baseline. This means that prior information is not very useful to predict commodity dynamics.

Table 3 Top 5 predicted directions and their proportion in the total export gain

No	Actual	ARIMA	Baseline model	Advanced model
1	Saudi Arabia	Libya	Azerbaijan	Italy
2	China	Spain	Georgia	Spain
3	Turkey	Ukraine	Uzbekistan	Libya
4	Azerbaijan	Kazakhstan	Ukraine	Ukraine
5	Italy	Georgia	China	Armenia
$	374755k	49666k	145263k	47982k
%	**79.3**	**13.6**	**31.8**	**13.1**

Table 3 presents five countries with the highest expected import growth from the Russian Federation. From this table we conclude that Russia export is not only commodity-non-diversified, but also partner-non-diversified. From this table we can see that purely prior-based "baseline" model performed best: it predicted more than 30% of actual export growth. ARIMA and "advanced" model performed approximately equally. So, we conclude that almost no new markets are explored: we will trade tomorrow with those, who we trade today. Additional unaccounted factors may include politics, wars, sanctions, etc.

7 Conclusion and future work

In this paper we have reviewed and discussed the problem of export growth points discovery. The main contribution of this paper is an automated data-driven framework that addresses the problem. The framework uses open data from many data sources and modern machine learning techniques. We also conducted preliminary experiments to evaluate the possibility to use retrospective data to rank growth point candidates. The experiments were based on open data from FAOSTAT and UN Comtrade.

Currently, it is very difficult to say for sure, which method is more useful for the final task – growth point discovery. Different methods compared to each other differently, depending on how to compare (top5 growth points, top5 commodities or top5 directions). This fact gives some clues on what a better model should look like. Another thing that has to be changes is the objective function: predicting short-term export value changes is very difficult and useless, because developing a new manufacture needs much more than one year. Thus, it makes much more sense to predict long-term trends.

Main directions of future work include (a) repeating experiments with adjusted methodology; (b) creating a manually-annotated dataset of growth points; (c) incorporating information about global economic situation and substitutes.

Acknowledgment

The research is supported by Russian Foundation for Basic Research, project 16-29-12877.

References

[1] Rodrik D. Institutions, integration, and geography: In search of the deep determinants of economic growth //In Search for Prosperity: Analytic Narratives on Economic Growth. Princeton University Press, Princeton. – 2003

[2] Food and Agriculture Organization of the United Nations. URL: http://www.fao.org/faostat/en/

[3] UN Comtrade: International Trade Statistics. URL: https://comtrade.un.org/data/

[4] Kaukin A., Idrisov G. The gravity model of Russia's international trade: the case of a large country with a long border. Working paper. – 2014

[5] Gordeev D. et al. Analysis of Global Supply Chains in International Trade Patterns. – 2016. – №. 765

[6] Duenas M., Fagiolo G. Modeling the International-Trade Network: a gravity approach //Journal of Economic Interaction and Coordination. – 2013. – Vol. 8(1). – pp. 155-178

[7] Jaud M., Kukenova M., Strieborny M. Financial Development and Sustainable Exports: Evidence from Firm product Data //The World Economy. – 2015. – Vol. 38(7). – pp. 1090-1114

[8] Snijders T. A. B. Models for longitudinal network data //Models and methods in social network analysis. – 2005. – Vol. 1. – pp. 215-247

[9] Grater S. et al. Linking export opportunities of products and services: the case of South Africa.

[10] Sgrignoli P. The World Trade Web: A Multiple-Network Perspective //arXiv preprint arXiv:1409.3799. – 2014

[11] Lall S., Weiss J., Zhang J. The "sophistication" of exports: a new trade measure //World Development. – 2006. – Vol. 34(2). – pp. 222-237.

[12] Bernard A. B., Jensen J. B. Why some firms export //Review of Economics and Statistics. – 2004. – Vol. 86(2). – p. 561-569

[13] Barigozzi M., Fagiolo G., Garlaschelli D. Multinetwork of international trade: A commodity-specific analysis //Physical Review E. – 2010. – Vol. 81(4). – p. 46-104

[14] Peluso S. et al. International Trade: a Reinforced Urn Network Model. – 2016. – №. 1601.03067.

[15] Fronczak A. Structural Hamiltonian of the international trade network //No. – 2012. – Vol. 1. – No. arXiv: 1205.4589. – pp. 31-46

[16] Shen B., Zhang J., Zheng Q. Exploring multi-layer flow network of international trade based on flow distances //arXiv preprint arXiv:1504.02361. – 2015

[17] Kelle M. et al. Cross border and Foreign Affiliate Sales of Services: Evidence from German Microdata //The World Economy. – 2013. – Vol. 36(11). – pp. 1373-1392

[18] Sujjaviriyasup T., Pitiruek K. Agricultural Product Fore-casting Using Machine Learning Approach //Int. Journal of Math. Analysis. – 2013. – Vol. 7. – №. 38. – p. 1869-1875

[19] Sujjaviriyasup T., Pitiruek K. Hybrid ARIMA-support vector machine model for agricultural production planning //Applied Mathematical Sciences. – 2013. – Vol. 7. – №. 57. – p. 2833-2840

[20] Microsoft. https://github.com/microsoft/lightgbm

Ключевой доклад 3

Keynote Talk 3

The EU's Human Brain Project (HBP) Flagship – Accelerating brain science discovery and collaboration

(Extended Abstract)

© Katrin Amunts,

Jülich, Düsseldorf, Germany

Abstract. The human brain has a multi-level organisation and high complexity. New approaches are necessary to decode the brain with its 86 billion nerve cells, which form extensive networks. Ultra-high resolution models of the brain pose massive challenges in terms of data processing, visualisation and analysis. The Human Brain Project creates a cutting-edge European infrastructure to enable cloud-based collaboration among researchers from different disciplines around the world. The infrastructure includes development platforms with databases, workflow systems, petabyte storage. New technologies, including neuromorphic computing and robotics are being developed. Neuroscientists and ICT-specialists collaborate in a co-design process, based on neuroscientific use cases, to develop such infrastructure, thus opening new perspective to decode the human brain.

Keywords: data analytics, data management, future computing, neuroinformatics, neuroscience.

1 Introduction

The human brain has a highly complex, multi-level organisation. New approaches are necessary to decode the brain with its 86 billion nerve cells, which form complex networks – a challenge that is addressed by the Human Brain Project (1). E.g., 3D Polarized Light Imaging (2) elucidates the connectional architecture of the brain at the level of axons, while keeping the topography of the whole organ; it results in data sets of several Petabytes per brain, which should be actively accessible while minimizing their transport. Thus, ultra-high-resolution models pose massive challenges in terms of data processing, visualisation and analysis.

High-resolution data obtained in post-mortem brains supplement information about structural and functional connectivity as obtained, e.g., by Magnetic Resonance imaging, acquired from the living human brain, but with lower spatial resolution. By bringing together different aspects of connectivity in one and the same atlas, it is feasible to combine the advantages of the different approaches, and to address the different spatial (and temporal) scales.

2 The Human Brain Project

The Human Brain Project is developing a new European research infrastructure. Part of it is the HBP atlas. It integrates information from multiple levels of brain organisation including cellular architecture, connectivity, molecular and genetic maps, as well results from neuroimaging and physiological studies. This atlas is a resource for empirical research, but also modeling and simulation. To represent microscopical data, it includes the Big Brain as one of its templates (3).

Proceedings of the XIX International Conference "Data Analytics and Management in Data Intensive Domains" (DAMDID/RCDL'2017), Moscow, Russia, October 10–13, 2017

The HBP infrastructure will enable cloud-based collaboration among researchers coming from different disciplines around the world. To achieve this, research platforms include databases, workflow systems, petabyte storage, and supercomputers, to address the requirements of the different users.

3 Project Structure

The core project involves 117 institutions and 500 researchers from 19 countries. It is organized in 12 Subprojects. Neuroscience Subprojects include: Mouse Brain Organization, Human Brain Organization, Systems and Cognitive Neuroscience and Theoretical Neuroscience. Activities in these Subprojects provide insights across all levels of brain organization: They will be more and more linked and integrated into comprehensive models, and tested by simulation.

Big data analytics and simulation are highly challenging and data intensive. Therefore, the Human Brain Project embraces ICT solutions. These include cloud-based collaboration and development platforms, with databases for metadata and provenance tracking, as well as data analytics and compute services, right up to leading-edge supercomputers, neuromorphic systems (4), and virtual robots (5). Research and development in these areas is organized in six Subprojects, forming research platforms: Neuroinformatics, Simulation, High Performance Analytics and Computing, Medical Informatics, Neuromorphic Computing, and Robotics. In addition, Management and Ethics & Society form two Subprojects.

The main IT service infrastructures is represented by the Neuroinformatics Platform (NIP) and the High-Performance Analytics and Computing Platform (HPAC). Data from patients are being collected and analysed in the Medical Informatics Platform (6). The

so-called COLLAB provides a connection between the Platforms, and a uniform entrance point. As a part of the NIP it serves as the main collaboration infrastructure.

The development of the platforms is guided by co-design projects where neuroscientific investigators and IT specialists collaborate. After the Ramp-up phase of the HBP from 2013 to 2016, the platforms were released in March 2016, can be accessed by test-users: https://www.humanbrainproject.eu/en/.

The NIP is conceptualized in such a way, that researchers from other countries can share their data, tools and expertise, to decode the human brain. This includes also collaboration with other brain initiatives (7).

4 Supercomputing

To obtain the means needed to address the incredible complexity of the brain, the neuroscience community will need to become a competitive player on high-end supercomputers and systems for big-data analytics. The HPAC aims to provide the Human Brain Project and in extension the neuroscience community with High Performance Computing systems geared towards the particular needs of neuroscientists.

Four European Tier 0 supercomputer centers currently are members of the consortium of the HBP: the Barcelona Supercomputing Centre (BSC) in Spain, the Consorzio Interuniversitario per le Applicazioni di Supercalcolo per Università e Ricerca (CINECA) in Italy, the Centro Svizzero di Calcolo Scientifico (CSCS) in Lugano, Switzerland, and the Jülich Supercomputing Centre (JSC) in Germany.

At the JSC, two new pilot systems for an interactive supercomputer have been obtained as the result of the "Pre-Commercial Procurement": JULIA, developed by Cray, and JURON from IBM and NVIDIA. These systems are specifically designed for data-intensive analytics and simulation applications in the neurosciences.

5 Teaching

The convergence of neuroscience and ICT that the Human Brain Project envisions in the long run depends on training a new generation of young researchers to be fluent in both areas. To this end, the HBP Education Programme offers online courses, advanced HBP Schools, an annual HBP Student Conference and Young Researchers Events.

All educational offers are made available through its website: https://education.humanbrainproject.eu/web/hbp-education-portal

References

[1] Amunts K, Ebell C, Muller J, Telefont M, Knoll A, Lippert T.: The Human Brain Project: Creating a European research infrastructure to decode the human brain. Neuron 2016; 92 (3): 574–81 doi: 10.1016/j.neuron.2016.10.046

[2] Axer M, Amunts K, Grässel D, Palm C, Dammers J, Axer H, Pietrzyk U, Zilles K.: A novel approach to the human connectome: Ultra-high resolution mapping of fiber tracts in the brain. Neuroimage 2011; 54: 1091–101. doi: 10.1016/j.neuroimage.2010.08.075

[3] Amunts K, Lepage C, Borgeat L, Mohlberg H, Dickscheid T, Rousseau MÉ, Bludau S, Bazin PL, Lewis LB, Oros-Peusquens AM, Shah NJ, Lippert T, Zilles K, Evans AC.: BigBrain – an ultra-high resolution 3D human brain model. Science 2013; 340: 1472–5 doi: 10.1126/science.1235381

[4] Friedmann S, Schemmel J, Grubel A, HArteil A, Hock M, Meier K.: Demonstrating hyrid learning in a flexible neuromorphic hardware system. IEEE Trans Biomed Circuits Syst 2017; 11 (1):128-142 doi: 10.1109/TBCAS.2016.2579164. Epub 2016 Sep 9

[5] Falotico E, Vannucci L, Ambrosano A, Albanese U, Ulbrich S, Vasquez Tieck JC, Hinkel G, Kaiser J, Peric I, Denninger O, Cauli N, Kirtay M, Roennau A, Klinker G, Von Arnim A, Guyot L, Peppicelli D, Martínez-Cañada P, Ros E, Maier P, Weber S, Huber M, Plecher D, Röhrbein F, Deser S, Roitberg A, van der Smagt P, Dillman R, Levi P, Laschi C, Knoll AC, Gewaltig MO.: Connecting artificial brains to robots in a comprehensive simulation framework: The Neurorobotics Platform. Front Neurorobot. 2017 Jan 25;11:2. doi: 10.3389/fnbot.2017.00002. eCollection 2017

[6] Frackowiak, R, Markram, H: The future of human cerebral cartography: a novel approach. Philos Trans R Soc Lond B Biol Sci. 2015 May 19; 370(1668): 20140171. doi: 10.1098/rstb.2014.0171

[7] Vogelstein, J. T. & Neuro Cloud Consortium: To the Cloud! A Grassroots Proposal to Accelerate Brain Science Discovery. Neuron. 2016 Nov 2;92(3):622-627. doi: 10.1016/j.neuron.2016.10.033

Проекты анализа данных в нейронауке

Data analysis projects in neuroscience

Search for Gender Difference in Functional Connectivity of Resting State fMRI

© Dmitry Kovalev[1] © Sergey Priimenko[2] © Natalya Ponomareva[3]

[1] Federal Research Center "Computer Science and Control" of Russian Academy of Sciences, Moscow, Russia
[2] Lomonosov Moscow State University, Moscow, Russia
[3] Research Center of Neurology, Moscow, Russia

dkovalev@ipiran.ru mior12@mail.ru ponomare@yandex.ru

Abstract. During past several year huge sets of fMRI data were obtained within Human Connectome Project. Despite this, technologies for scalable analysis of large amounts of data are rarely used to analyze whole data set. Authors conducted virtual experiment on a large sample of data taken from the HCP to find the gender differences in functional connectivity. A review of methods for search for the functional connectivity is fulfilled. Further analysis of distributed use and scalability on large datasets of rfMRI data is provided with the discussion of existing libraries and suggestions of how to integrate them with a distributed system. As a result, the distributed architecture of the software based on the Apache Spark framework is developed. Being fairly complex, it includes ontology, conceptual schema and workflow. The results of this experiment may be of interest to neurophysiologists for further analysis.

Keywords: data intensive research, distributed infrastructure, problem solving in neurophysiology.

1 Introduction

Today in many branches of science it is necessary to solve problems associated with increasing scale of data [1–3]. This led to the development of specialized tools, which primarily focus on structured data, but are increasingly being adapted for more general forms [4, 5]. Yet this tools and software are not widely used in data intensive research and methodology to correctly apply them has still to be developed. Different use-cases from multidisciplinary fields can greatly impact the evolution of this methodology and tools.

One of the most prominent examples of data intensive domains is the field of neurophysiology, where the amount of data has reached petabyte scale. Neurophysiology allows to visualize the structure, functions and biochemical characteristics of the brain. In particular, approaches to find the functional connections of the brain departments are being explored [6]. One way to do that is to measure the functional connectivity between brain regions as the level of co-activation of spontaneous functional time series of resting-state fMRI [7–9].

During past several years, major projects such as the Human Connectome Project (HCP) and the 1000 connectome have started with more than a thousand people participating. Datasets are open to the scientific community. Such large-scale data warehouses could serve as the beginning for the use of technologies for analyzing large amounts of data in the neuroimaging of the human brain, yet there are some limitations. One of the reasons why the community of neurobiologists do not use tools to work with large amounts of data is that standard file formats, such as NIFTI[10], are binary and possess additional costs to deliver to distributed file systems. Another problem is that many distributed systems do not effectively perform iterative algorithms, such as principal component analysis (PCA) and the independent component analysis (ICA), which are actively used in the field of neuroimaging.

One of significant are of research in neurophysiology is the study of gender difference in functional connectivity [11]. For example, there is a study of army veterans that experience physical and psychiatric complications, including craniocerebral trauma, post-traumatic stress and depression. The integration of a large number of women into military operations attracted attention to the potential sexual differences in the frequency and recovery from craniocerebral trauma, as well as from other concomitant disorders. Understanding the role of gender-related effects can provide information on the needs for evaluating treatment for women, which can demonstrate both similarities and differences from men.

This article aims at developing approach for a distributed analysis of data intensive neurophysiology domain. The article is structured as follows. Section 2 surveys existing distributed methods ant tools to process and analyze neurophysiological datasets. Section 3

Proceedings of the XIX International Conference "Data Analytics and Management in Data Intensive Domains" (DAMDID/RCDL'2017), Moscow, Russia, October 10–13, 2017

presents domain ontology that was created to better interact with domain experts. Section 4 describes distributed programming implementation on the existing computational infrastructure, as well as output results. Section 5 concludes the article.

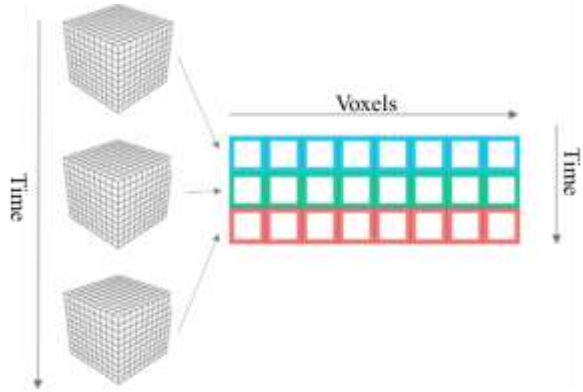


Figure 1 Transformation of 4-D array into 2-D array

2 Data analysis methods

2.1 Data processing

Resting-state fMRI dataset from the HCP project is used. The HCP consortium has developed an information platform for storing raw and processed data, systematic processing and analysis of data, obtaining and researching data. One of the main components of the project is ConnectomeDB. ConnectomeDB provides database services for the storage and dissemination of datasets that are open to the scientific community. The data is already preprocessed. Preprocessing consists of removing spatial artifacts, distortion, surface formation and alignment to a single standard space.

Data processing is divided into two parts: data cleaning inside the brain (FMRIVolume) and on the brain surface (FMRISurface) [3].

At the FMRIVolume stage, spatial distortion removal, volume redistribution due to subject movement during the session, normalization of 4D images to the standard value and creation of the final brain mask are done.

The main purpose of FMRISurface is to display time series in the standard CIFTI space. This is achieved by comparing the voxels in the cortical region of the gray matter to the native surface of the cortex and transforming each subcortical region for each individual to a standard set of voxels for each data set.

After processing the data, resting-state fMRI time series are stored in a special format – NIFTI. As a result, the data obtained with the resting-state fMRI yields more than 10 TB obtained for more than 1000 people. During the experiment, each patient was placed in a dark room and asked to relax, but not to fall asleep. The experiment was conducted in 4 sessions for 15 minutes. Two sessions of the fMRI device took pictures from the left side of the brain to the right side of the brain, and the other two sessions from the right side of the brain to the left.

2.2 Data analysis methods

The data of each subject is represented as a matrix $X_{(txv)}$ (see Fig. 1), where each row represents a set of voxels of the brain at a particular time, and each column is a time series for the corresponding voxel [12]. It is assumed that the data has already been pre-processed to remove artifacts and scaled to a standard space (coordinate system) so that the voxels are anatomically compatible for all subjects. It is also assumed that the time series of each voxel is shifted by its mean (and, possibly, normalized to the variance) [13].

If the data set consists of one object, in order to reduce the dimensionality of the data, the PCA is applied:

$$X_{(txv)} = U_{(txn)} \times S_{(nxn)} \times V^T_{(nxv)},$$

where n is the number of main components (usually much smaller than t), U is the set of temporal eigenvectors, V is the set of spatial eigenvectors, and the corresponding eigenvalues on the main diagonal of the matrix S (n largest eigenvalues). Then, ICA is applied to the matrix V, estimating a new set of spatial components that are linear combinations of the vectors of the matrix V and are maximally independent of each other. If the data set consists of several subjects, then initially all the data is combined into one large set consisting of s subjects, and then PCA and ICA are applied. The resulting approximation will be the same as above, but now with dimensions $n * s \times t$ (see Fig. 2).

With large data sets, or with a large number of subjects, it becomes unreasonable to form a complete set of data, and then apply PCA and ICA due to memory and time limitations. To solve this problem, several algorithms were invented.

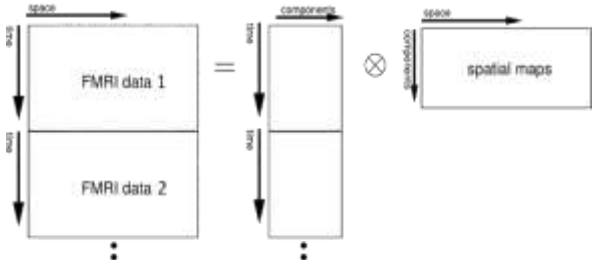


Figure 2 PCA for concatenated data

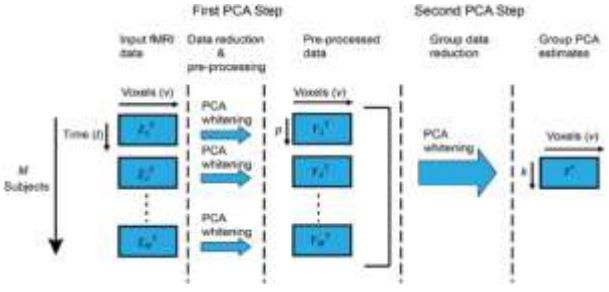


Figure 3 Parallel execution of PCA

In 2001, it was suggested to approximate the concatenation of all data sets by first reducing each set of data to m main spatial vectors using PCA and then concatenating them and applying the final PCA to reduce

the final dataset to n components and then apply ICA [14]. Although using a small value of m limits the memory requirements for these operations, the data size is scaled linearly with the number of objects, which can eventually become impractically large. In addition, an important piece of information can be lost if m is not relatively large (usually it should not be large). Information can be difficult to assess at the level of an individual subject, but it can be important at the group level (see Fig. 3).

To overcome these limitations, the MELODIC's Incremental Group-PCA (MIGP) algorithm was proposed[15]. MIGP is an incremental approach, the goal of which is to provide a very close approximation to the complete concatenation of the data set followed by the PCA, but without large memory requirements. High accuracy is achieved due to the fact that individual sets of subjects' data do not decrease to a small number of components of PCA. The incremental approach preserves the inner space of PCA from m weighted spatial eigenvectors, where m is usually larger than the number of time points in each individual data set. By "weighted" is meant that the eigenvalues are included in the matrix of spatial eigenvectors. The final set of m components representing the temporarily concatenated output of the PCA can then be reduced to the required dimension n simply by storing the upper n components and, if necessary, discarding the weighting coefficients (eigenvalues).

Usually, 2–3 sets of data are first concatenated. This data set is then fed into an m-dimensional PCA and following matrix is obtained: $W_{(m\times v)} = S_{(m\times m)} \times V^{T}_{(m\times v)}$. Each vector is multiplied by its own value. The eigenvalues characterize the importance of the component here, so statistical information is not lost. W becomes the current evaluation of the group set and can be considered as a matrix of pseudo-series consisting of m time points and v voxels. For each data set of each subject, we gradually update W by combining W with each data set X_i and applying the ICA to get the updated W , saving only m main components. Thus, the variance of each batch of data is preserved (see Fig. 4).

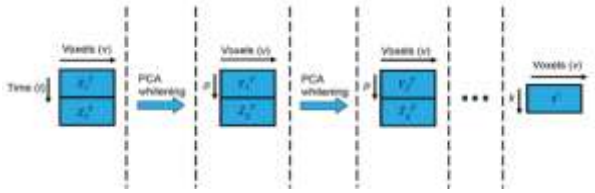


Figure 4 MELODIC Incremental Group PCA

MIGP does not increase the memory requirement with an increase in the number of subjects, large matrices are never formed, and the computation time varies linearly with the number of objects. This is easily parallelized by applying the approach in parallel to subsets of entities, and then combining them using the same approach of "concatenation and reduction" described above.

2.3 Libraries

Nibabel [16] is a library that provides an API for reading and writing some common file formats for neuroimaging. These formats include: ANALYZE (plain, SPM99, SPM2 and higher), GIFTI, NIfTI1, NIfTI2, MINC1, MINC2, MGH and ECAT, as well as Philips PAR/REC. Different image format classes provide full or selective access to header information (meta), and access to image data is made available through the arrays of the numpy library.

Objects of the image of nibabel consist of three elements:

1. The *n*-dimensional array containing the image
2. Matrix of affine transformations of size 4x4, which correlates the image coordinates with the standard world coordinate space.
3. Image metadata, stored in the header.

When an image is loaded, an object of type Nifti1mage is created. The file name can have an extension of both .nii and .nii.gz.

It is worth noting that when the load function is called directly, image data is not loaded into memory, since images can be stored as a numpy array or stored on a disk. To load data from a disk, you need to call the *get_data()* function of an object of type Nifti1Image. This function returns an n-dimensional numpy array.

In addition, an object of type Nifti1Image is created from numpy arrays. To do this, one should pass an n-dimensional data array and an affine transformation matrix to the Nifti1Image constructor must.

Nitime[17] is a library for the analysis of time series in the field of neuroimaging. Nitime can be used to represent, process and analyze time series data from experimental data. The main purpose of the library is to serve as a platform for analyzing data collected in neurophysical experiments. The basic principle of nitime implementation is the division of time series representation and time series analysis.

An important feature of the nitime library is lazy initialization. Most attributes of both time series and analysis objects are used only when necessary. That is, the initialization of a time series object or an analysis object does not cause any intensive calculations. In addition, after the calculation starts, the object is saved and ensures that access to the results of the analysis will cause the calculation to be performed only when the analysis is performed for the first time. After that, the result of the analysis is saved for further use.

One of the algorithms of the nitime library is the correlation analysis of brain regions. It calculates the correlation between one time series that represents a given area of the brain, with other areas that are also represented by a time series. To calculate the correlation between regions in the nitime library, there is a SeedCoherencAnalyzer function that takes two time series inputs and returns a correlation matrix that can be used for further analysis.

Nilearn [18] is a Python module for statistical

processing of neuroimaging data.

It uses scikit-learn module for multidimensional statistics with applications in intelligent modeling, classification, decoding, and connectivity analysis. Nilearn can work NiftiImage objects from the nibabel library.

Nilearn library has great functionality for working with nii-images. It allows visualizing, decoding, exploring the functional connectivity, and performing various manipulations, such as smoothing, marking and advanced statistical analysis.

Nilearn provide CanICA method that is the ICA method for analyzing fMRI data at the group level. Compared to other strategies, it brings a well-controlled group model, as well as a threshold algorithm that controls specificity and sensitivity with an explicit signal model.

In order to get a time series and build a correlation matrix for it, nilearn provides the NiftiMapsMasker object. To create an object, one needs to specify an atlas of the brain regions.Nilearn provides the ability to create a correlation matrix for independent components that iscomputed by CanICA.

3 Ontology

The study of neuroimaging with large amounts of data represents the intersection of different areas of science. In order to use the same terms and concepts, simple ontology was developed that describes the main entities used in this work and a conceptual schema that defines the types of data, constraints on these data types and the means of interaction between them. Ontology is a formal specification of shared conceptualization [19].

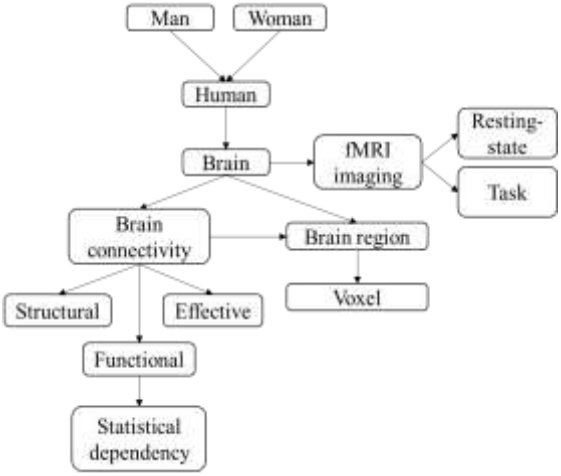


Figure 5 Main concepts of the domain ontology

The ontological specification of the subject area of neuroimaging consists of the following components (see Fig. 5):

- Neuro-image – a 3-dimensional or 4-dimensional image (a series of 3-dimensional images), reflecting the distribution of metabolic activity in different regions of the brain in different time intervals [20].
- The area of the brain is a set of voxels, sorted by a certain feature. Most often presented in the form of time series [20].
- Voxel is an element of a three-dimensional image containing some value.
- Independent models - a model for investigating thefunctional connectivity of the entire brain. They are designed to search for general patterns of functional connectivity between brain regions.Dependent models are a model for analyzing the correlation of a given region of the brain.
- Brain connectivity – the structure of anatomical connections, statistical dependencies or cause-effect interactions between individual units within the brain's nervous system [21].
- Structural connectivity refers to a network of physical or structural links linking sets of neurons or neural elements to structural biophysical features [22].
- Functional connectivity is a statistical type of connection between anatomically unconnected areas of the brain that have common functional properties [7].
- Effective connectivity – the combination of structural and functional connectivity. It describes the networks of directions of one neural element over another.
- The resting-state fMRI is a neural image obtained as a result of an experiment when the subject was at rest and did not engage in active tasks.
- The task fMRI is the neuro-images obtained as a result of the experiment, when the subject performed active actions, e. g., listened to music.

4 Implementation

4.1 Laboratory cluster specifications

Virtual experiment was executed on the laboratory cluster (see Fig. 6). It consists of 2 master nodes and 6 slave nodes. Each master node has 32Gb of RAM, 24 threads and 2 Tb of disk space in RAID1. Slave nodes have 64Gb of RAM, 24 threads and 4 Tb of disk space attaches as JBOD. All the machines are connected to 10Gbs switch.

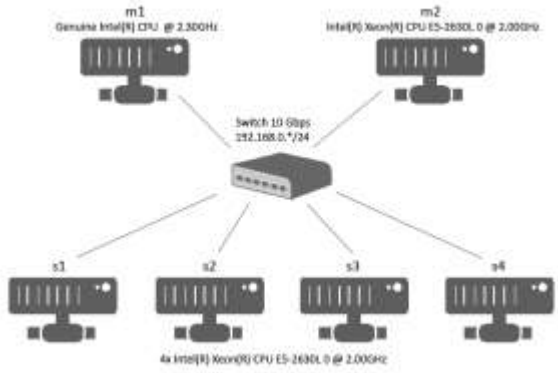


Figure 6 Cluster Architecture

On the cluster, the Hortonworks Data Platform (HDP) distribution package is installed. This distribution

represents a set of tools from the Hadoop infrastructure running Apache Ambari.

A distributed file system (HDFS) (Hadoop Distributed File System) is installed file system. HDFS consists of a NameNode server and DataNode servers. The NameNode server manages the namespace of the file system and manages the clients' access to the data. The main NameNode server is installed on the *m1*node and records all transactions associated with changing the file system metadata to a special file called *EditLog*. When rt the main NameNode server is started, it reads the HDFS image and applies all the changes to it. This is done once at startup. A similar operation is performed by the Secondary NameNode, which is installed on the *m2* machine. On machines *s1-s4* DataNode servers are installed, which are responsible for storing the data itself and keeping its integrity.

For the sharing, scalability and reliability of the Hadoop cluster, a resource manager YARN [5] is used. YARN offers a hierarchical approach to the cluster infrastructure. The root of the YARN hierarchy is the ResourceManager. This daemon manages the entire cluster and assigns applications to the underlying computing resources. It allocates resources (computing resources, memory, and bandwidth) for the basic NodeManager. ResourceManager interacts with ApplicationMaster when allocating resources and with NodeManager when starting and monitoring basic applications. ResoureManager is located on *m2*, and NodeManager on nodes *s1–s4*.

Another important module for the Hadoop cluster is the Zookeper. ZooKeeper is a server that coordinates distributed processing. It provides a distributed configuration service, a synchronization service, and a registry of names for distributed systems. Distributed applications use ZooKeeper to store and notify updates of important configuration information. The Zookeper server is running on the *m1* node.

Since most of the calculations are iterative algorithms, Apache Spark was chosen as the computational backbone. Apache Spark provides a fast and versatile platform for data processing. In comparison with Hadoop, Spark accelerates the work of programs by minimizing disk input-output operations.

In Spark, the concept of RDD (stable distributed data set) is introduced – an unchangeable fault-tolerant distributed collection of objects that can be processed in parallel. RDD can contain objects of any type. RDD is created by loading an external data set or distributing a collection from the main program (driver program). In RDD, two types of operations are supported:

- Transformations are operations (for example, mapping, filtering, merging, etc.) performed over RDD. The result of the transformation is a new RDD containing its result.
- Actions are operations (eg, reduction, count, etc.) that return a value that results from some calculations in RDD.

The cluster has Spark History Server installed on *m1*, Spark Thrift Server on *m2*, Livy Server on *m2* and Spark Clients on all nodes.

For more convenient programming on a cluster, we use Apache Zeppelin – a web-based notebook that allows to conduct interactive data analytics. It supports many interpreters, including the Spark interpreter and the Python interpreter.

Scalability. As of algorithm used, each slave machine handles several independent fMRI images, so scalability increases almost linearly with using more slave nodes. It is bounded by the network speed when transmitting initial image data into slave memory, however the transmission time is several seconds and is negligible compared to processing time.

4.2 Workflow

Workflow is depicted on Fig. 7. The program reads all files from the directory, checks the validity of the format (all data are compressed zip folders). After that, the subject number is extracted from the file name and its gender is checked using an additional metadata file. When the gender is known, the file is unzipped to the corresponding folder. Inside the unzipped folder is a 4-D image in the .nii.gz format. Using the *nibabel* library, the image is loaded into memory as an array of type *numpy.array*. From this array, a new array is created with information about the spatial coordinates before the value of the voxel. The new array is compressed by the gzip algorithm and stored in HDFS.

Due to Apache Spark limitations files larger than 2.5 GB in binary format can not be loaded. In the uncompressed form, the sizeis 4.3 GB, so file needs to be compression. After compression, the file occupies just 700 MB.

Spark task is started with the following parameters:

- *num-executor=4* – number of executable entities;
- *executor-memory*=25 GB – the amount of memory used for one execution process;
- *executor-cores*=2 – the number of cores used for each executive entity.
- *driver-memory*=8 GB – the amount of memory used for the driver process, that is, where SparkContext is initialized.

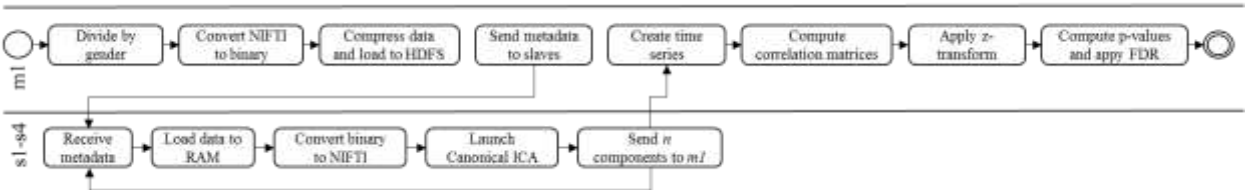


Figure 7 Workflow

YARN creates on each node a container that receives information from the driver. All calculations occur in two streams. When metadata is received, a file with a compressed binary array is loaded into memory. The program decompresses it and converts it to a normal array without information about the indexes. Then, using the resulting array and affinity transformation matrix, Nifti1Image and the CanICA object are created with the following parameters: *n_components*=20*; Smoothing_fwhm*=6*; N_init*=10*; Threshold*=3*; Verbose*=10.

The CanICA object is passed to the Nifti1Image object and an image consisting of 20 components is output. This image is returned to the m1 driver. Thus, each node receives a portion of the paths to the compressed images, processes them, and returns the result to the driver. The task is executed until all the files specified for analysis on the m1 driver are processed. When the nodes complete the tasks, the driver comes with a list of Nibabel1Image objects that contain independent components. The data for all objects is averaged and a time series is created using the NiftiLabelsMasker object.

A map of regions of the brain is transferred to the constructor of the NiftiLabelsMasker object. Using the ConnectivityMeasure object, which is created with the correlation parameter, the correlation matrix for the brain regions is considered.The correlation matrix for men and women is calculated separately. After this, the Fisher transform (z-transform) is applied to each matrix.

After a new sample is calculated, which is obtained as the difference between the male z_m obtained and the female sample z_w. This sample will have a normal

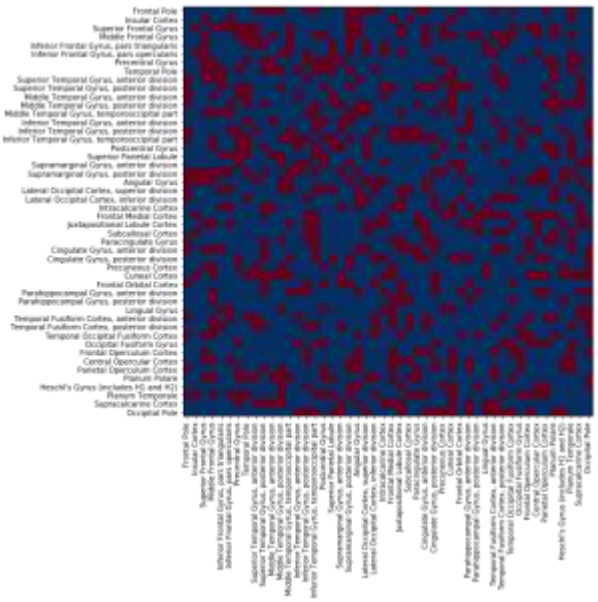


Figure 8 Binary matrix of functional connectivity difference

distribution with a mathematical expectation of 0 and a variance of 2/(n-3). For this sample, calculates a critical area with a significance level of 0.05 and c is corrected for multiple testing of the Benjamin–Hochberg hypotheses. As a result, a binary matrix is obtained that shows the deviation or acceptance of hypotheses for each area of the brain.

4.3 Results

In total 50 male and 50 female subjects are used. All data is resting-state fMRI images.

Fig. 8 depicts a binary matrix of gender differences in the functional connectivity of healthy middle-aged people. Red spots mark areas that correlate both in men and women, and blue dots indicate a lack of correlation. For example, this experiment shows that the upper front (Superior Frontal Gyrus) of the brain has a significant correlation with the insular cortex (Insular Cortex), but does not have a significant correlation with the front part (Frontal Pole) of the brain.

The independent components of averaged male subject show a greater functional connectivity compared to women. It can be seen that the main activity of the brain of men and women occurs near its cortex.

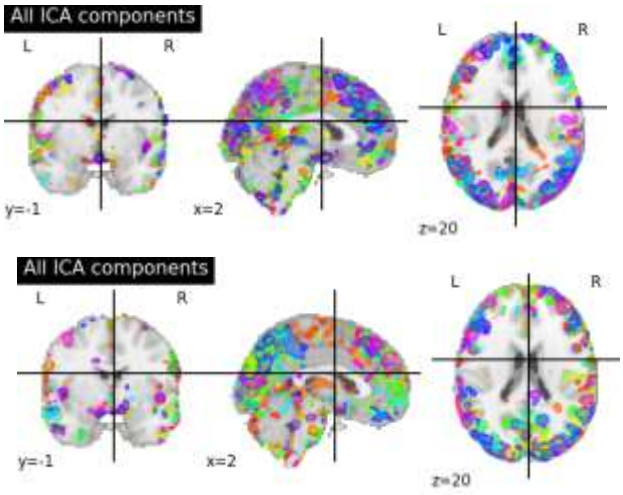


Figure 9 Averaged independent components for men (upper) and women (lower)

5 Conclusion

This paper presents distributed methods and means for searching gender differences in functional connectivity of resting-state fMRI were explored. Several methods for the search for functional connectivity of functionally magnetic resonance tomography of human rest are considered. To work with large amounts of data, machine learning methods were used to identify repetitive patterns and to intelligently reduce data. Their possibilities of parallel and distributed use and scaling are investigated with large amounts of input data. For the sake of better communication with domain experts the domain ontology was specified with main entities that describe this area and the necessary links between them.

The review of existing means of preparation and preprocessing of data on local and distributed systems is carried out. At the moment there are few libraries for working with the NIFTI format on a distributed system, so the input and output procedures for data were

implemented in this work. To preprocess the data, we used method compositions from the nibabel and nilearn libraries.To solve the problem, an overview of existing distributed systems was made, among which the Apache Spark framework was most effective. For the experiment, a cluster of 6 machines was taken, where the two machines were the main nodes, and 4 the workers. On the cluster, the minimum set of programs required for the experiment, such as YARN, HDFS, ZooKeeper, Spark and Zeppelin notebook was installed and configured.

A virtual experiment was performed in a distributed system. The time of this experiment was 4 hours for 400 GB of data. As a result of the experiment, matrices of connectivity between the brain regions of men and women were obtained, as well as a binary matrix of gender differences in functional connectivity.

Acknowledgments

This research was partially supported by the Russian Foundation for Basic Research (projects 15-29-06045, 16-07-01028).

References

[1] Council, N.R.: Frontiers in Massive Data Analysis. The National Academies Press, Washington, DC (2013)

[2] Hey, A.J., Tansley, S., Tolle, K.M., others eds: The Fourth Paradigm: Data-Intensive Scientific Discovery. Microsoft Research Redmond, WA (2009)

[3] Van Essen, D.C., Smith, S.M., Barch, D.M., Behrens, T.E., Yacoub, E., Ugurbil, K., Consortium, W.-M.H., others: The WU-Minn Human Connectome Project: An Overview. Neuroimage. 80, 62–79 (2013)

[4] Zaharia, M., Xin, R.S., Wendell, P., Das, T., Armbrust, M., Dave, A., Meng, X., Rosen, J., Venkataraman, S., Franklin, M.J., others: Apache Spark: A Unified Engine for Big Data Processing. Communications of the ACM, 59, pp. 56-65 (2016)

[5] Vavilapalli, V.K., Murthy, A.C., Douglas, C., Agarwal, S., Konar, M., Evans, R., Graves, T., Lowe, J., Shah, H., Seth, S., others: Apache Hadoop yarn: Yet Another Resource Negotiator. In: Proc. of the 4th annual Symposium on Cloud Computing. p. 5. ACM (2013)

[6] Huth, A.G., Heer, W.A. de, Griffiths, T.L., Theunissen, F.E., Gallant, J.L.: Natural Speech Reveals the Semantic Maps that Tile Human Cerebral Cortex. Nature, 532, pp. 453-458 (2016)

[7] Friston, K.J.: Functional and Effective Connectivity: A Review. Brain connectivity, 1, pp. 13-36 (2011)

[8] Biswal, B.B., Kylen, J.V., Hyde, J.S.: Simultaneous Assessment of Flow and BOLD Signals in Resting-state Functional Connectivity Maps. NMR in Biomedicine, 10, pp. 165-170 (1997)

[9] Biswal, B.B., Mennes, M., Zuo, X.-N., Gohel, S., Kelly, C., Smith, S.M., Beckmann, C.F., Adelstein, J.S., Buckner, R.L., Colcombe, S., others: Toward Discovery Science of Human Brain Function. Proc. of the National Academy of Sciences, 107, pp. 4734-4739 (2010)

[10] Cox, R.W., Ashburner, J., Breman, H., Fissell, K., Haselgrove, C., Holmes, C.J., Lancaster, J.L., Rex, D.E., Smith, S.M., Woodward, J.B., others: A (sort of) New Image Data Format Standard: Nifti-1. Neuroimage, 22, e1440 (2004)

[11] McGlade, E., Rogowska, J., Yurgelun-Todd, D.: Sex Differences in Orbitofrontal Connectivity in Male and Female Veterans With TBI. Brain imaging and Behavior, 9, pp. 535-549 (2015)

[12] Smith, S.M., Hyvärinen, A., Varoquaux, G., Miller, K.L., Beckmann, C.F.: Group-PCA for Very Large fMRI Datasets. NeuroImage, 101, pp. 738-749 (2014)

[13] Beckmann, C.F., Smith, S.M.: Probabilistic Independent Component Analysis for Functional Magnetic Resonance Imaging. IEEE Transactions on Medical Imaging, 23, pp. 137-152 (2004)

[14] Calhoun, V.D., Adali, T., Pearlson, G.D., Pekar, J.: A Method for Making Group Inferences from Functional MRI Data Using Independent Component Analysis. Human Brain Mapping, 14, pp. 140-151 (2001)

[15] Rachakonda, S., Silva, R.F., Liu, J., Calhoun, V.D.: Memory Efficient PCA Methods for Large Group ICA. Frontiers in Neuroscience, 10 (2016)

[16] Gorgolewski, K., Burns, C.D., Madison, C., Clark, D., Halchenko, Y.O., Waskom, M.L., Ghosh, S.S.: Nipype: A Flexible, Lightweight and Extensible Neuroimaging Data Processing Framework in Python. Frontiers in Neuroinformatics, 5 (2011)

[17] Rokem, A., Trumpis, M., Perez, F.: Nitime: Time-Series Analysis for Neuroimaging Data. In: Proc. of the 8th Python in Science Conf., pp. 68-75 (2009)

[18] Abraham, A., Pedregosa, F., Eickenberg, M., Gervais, P., Mueller, A., Kossaifi, J., Gramfort, A., Thirion, B., Varoquaux, G.: Machine Learning for Neuroimaging With Scikit-learn. Frontiers in Neuroinformatics, 8 (2014)

[19] Sowa, J.F., others: Knowledge Representation: Logical, Philosophical, and Computational Foundations. MIT Press (2000)

[20] Poldrack, R.A.: Region of Interest Analysis for fMRI. Social Cognitive and Affective Neuroscience, 2, pp. 67-70 (2007)

[21] Van Den Heuvel, M.P., Pol, H.E.H.: Exploring the Brain Network: A Review on Resting-State fMRI Functional Connectivity. European Neuropsychopharmacology, 20, pp. 519-534 (2010)

[22] Sporns, O.: Discovering the Human Connectome. MIT Press (2012)

Исследование методов организации виртуального эксперимента для задачи поиска эффективной связности функциональной магнитно-резонансной томографии действия человека

© Д.С. Ендеева

Московский государственный университет имени М.В. Ломоносова,
Москва, Россия

endeevavmk@gmail

Аннотация. Быстрое развитие информационных технологий позволяет резко увеличить как объем накопленных данных, так и способствует продвижению новых методов анализа этих данных. Работа посвящена исследованию техники проведения виртуального эксперимента на примере задачи анализа когнитивных функций человеческого мозга. Задача происходит из одной из самых динамично развивающихся областей с интенсивным использованием данных – нейрофизиологии. В процессе исследования построена онтология предметной области, описаны характерные особенности данных, явно специфицированы используемые априорные гипотезы. На примере одного из исследований формально описан поток работ поиска эффективной связности данных функциональной магнитно-резонансной томографии человека. Предполагается, что формальное описание всех этапов проведения виртуального эксперимента может быть полезным как исследователям из названной области, так и при проведении анализа данных из других мультидисциплинарных областей.

Ключевые слова: аналитика и управление данными, интенсивное использование данных, виртуальный эксперимент, когнитивные функции мозга.

Organizing a Virtual Experiment for the Analysis of Effective Connectivity of Human Task Functional Magnetic Resonance Imaging

© Darya Endeeva

Lomonosov Moscow State University,
Moscow, Russia

endeevavmk@gmail.com

Abstract. The rapid development of information technologies allows us to dramatically increase both the volume of accumulated data and promote new methods for analyzing this data. The article is devoted to the study of the technique of managing virtual experiment with the example of the problem of analyzing the cognitive functions of the human brain. The task comes from one of the most dynamically developing areas with the intensive use of data – neurophysiology. In the process of research, the ontology of the domain is constructed, the characteristic features of the data are described, the a priori hypotheses used are explicitly specified. Formally specified workflow for searching the effective connectivity of human fMRI is extracted. It is assumed that a formal description of all stages of the virtual experiment can be useful both for researchers from the given area and for analyzing data from other multidisciplinary areas.

Keywords: analytics and data management, data intensive domains, virtual experiment, cognitive functions.

1 Введение

Развитие информационных технологий оказывает огромное влияние на экспериментальную, теоретическую и вычислительную науку. Развитие технологий сильным образом повлияло на нейроинформатику, в которой происходит накопление огромного массива данных, развиваются новые методы, требующие больших

Труды XIX Международной конференции «Аналитика и управление данными в областях с интенсивным использованием данных» (DAMDID/ RCDL'2017), Москва, Россия, 10–13 октября 2017 года

вычислительных мощностей. Данные, в свою очередь, становятся основным источником новых научных открытий. Примерами крупных проектов в области нейрофизиологии являются проекты Human Connectome Project (HCP), NITRC: 1000 Functional Connectomes. Объем накопленных данных в этих проектах достигает петабайтов. Например, проект HCP содержит данные функциональной магнитно-резонансной томографии (фМРТ) около 1200 человек, полученные на сканерах 3T MR и 7T MR, а также данные магнитоэнцефалографии (МЭГ) 95 человек (среди которых структурные изображения; изображения, полученные в состоянии покоя субъекта; изображения, полученные при выполнении субъектом заданий; диффузные данные).

Появляется потребность в семантической методологии, которая упрощает моделирование научных знаний, проверку гипотез, семантическую интеграцию данных, анализ данных для различных дисциплин для использования неспециалистами. Важной задачей исследователей является обеспечение повторного использования методов другими исследователями и воспроизводимости результатов эксперимента, для чего процессы работ должны снабжаться метаданными и спецификациями методов.

Для организации обработки данных и систематического, повторяемого и воспроизводимого выполнения всех процедур, являющихся частью процесса исследования, становится целесообразно описывать и разрабатывать потоки работ, которые представляют собой точное описание научной процедуры, состоящей из множества последовательных шагов. Продуманные повторения потоков работ автоматически дают новые результаты при доступности новых исходных данных и новых результатов, а также новых методов. Сами рабочие процессы, как важная часть науки, ориентированной на данные, могут создаваться и динамически трансформироваться, исходя из текущих потребностей. Объединение рабочих процессов с опубликованными результатами обеспечивает прозрачность и сравнимость исследований, что может ускорить процесс научных открытий.

Целью работы является исследование методов и средств для организации виртуальных экспериментов по анализу эффективной связности функциональной магнитно-резонансной томографии действия здоровых людей. Достижение цели предполагает решение следующих задач: разработка онтологии и концептуальной схемы предметной области, которая требуется для понимания процессов, происходящих в мозге, и последующего моделирования этих процессов; формирования гипотез; построение виртуального эксперимента; формализация потока работ виртуального эксперимента; обзор существующих решений по обработке и анализу данных функциональной магнитно-резонансной томографии, требующихся на разных этапах потока работ; исследование методов анализа эффективной связности функциональной магнитно-резонансной томографии. Под виртуальным экспериментом подразумевается относительно новое направление в научно-исследовательском процессе, обусловленное реализацией моделей средствами вычислительной техники.

Основой для построения потока работ послужил виртуальный эксперимент, посвященный поиску эффективной связности между двумя областями мозга – V5 и pSTS – на основе фМРТ данных на задаче восприятия одушевленного движения методом Динамического Каузального Моделирования (Dynamic Causal Modelling, DCM) [1].

Сенсорная зрительная зона V5 (или MT, от англ. Middle Temporal) считается чувствительной к любому типу визуального движения [3]. Регион pSTS (от англ. posterior superior temporal sulcus) – задняя область верхней височной борозды – во многих исследованиях (например, в [4]) связывается с социальной перцепцией, анализом и восприятием биологических движений и анализом намерений человеческих субъектов.

Участникам показывали серию коротких видеоклипов, в которых простые фигуры (треугольники, квадраты, круги) либо взаимодействовали друг с другом (одушевленное движение), либо двигались механически (неодушевленное движение). Одушевленное движение обычно более сложное, нелинейное и, таким образом, менее предсказуемое по сравнению с неодушевленным движением, так как основано на скрытых намерениях агента, производящего движение. Неодушевленное движение, наоборот, происходит благодаря действиям относительно неизменных сил (физических законов) и поэтому считается более предсказуемым. Умение предсказывать одушевленное движение является ключевым для выживания, что подчеркивает актуальность исследования восприятия движения.

2 Задача исследования когнитивных функций мозга

2.1 Онтология предметной области

Для корректной совместной работы с экспертами из области была разработана онтология предметной области (см. Рис. 1).

С помощью визуализации методом магнитно-резонансной томографии (МРТ) можно изучать структуру (анатомию) мозга, а с помощью визуализации методом функциональной магнитно-резонансной томографии – его функциональность. Изображения, полученные методами МРТ и фМРТ, собирают с одного сканера. Единицей объёма получаемого со сканера трехмерного изображения является воксель. Изображение, полученное методом МРТ, представляет собой трехмерный массив вокселей. Слайсом называется двумерный массив вокселей нейроизображения.

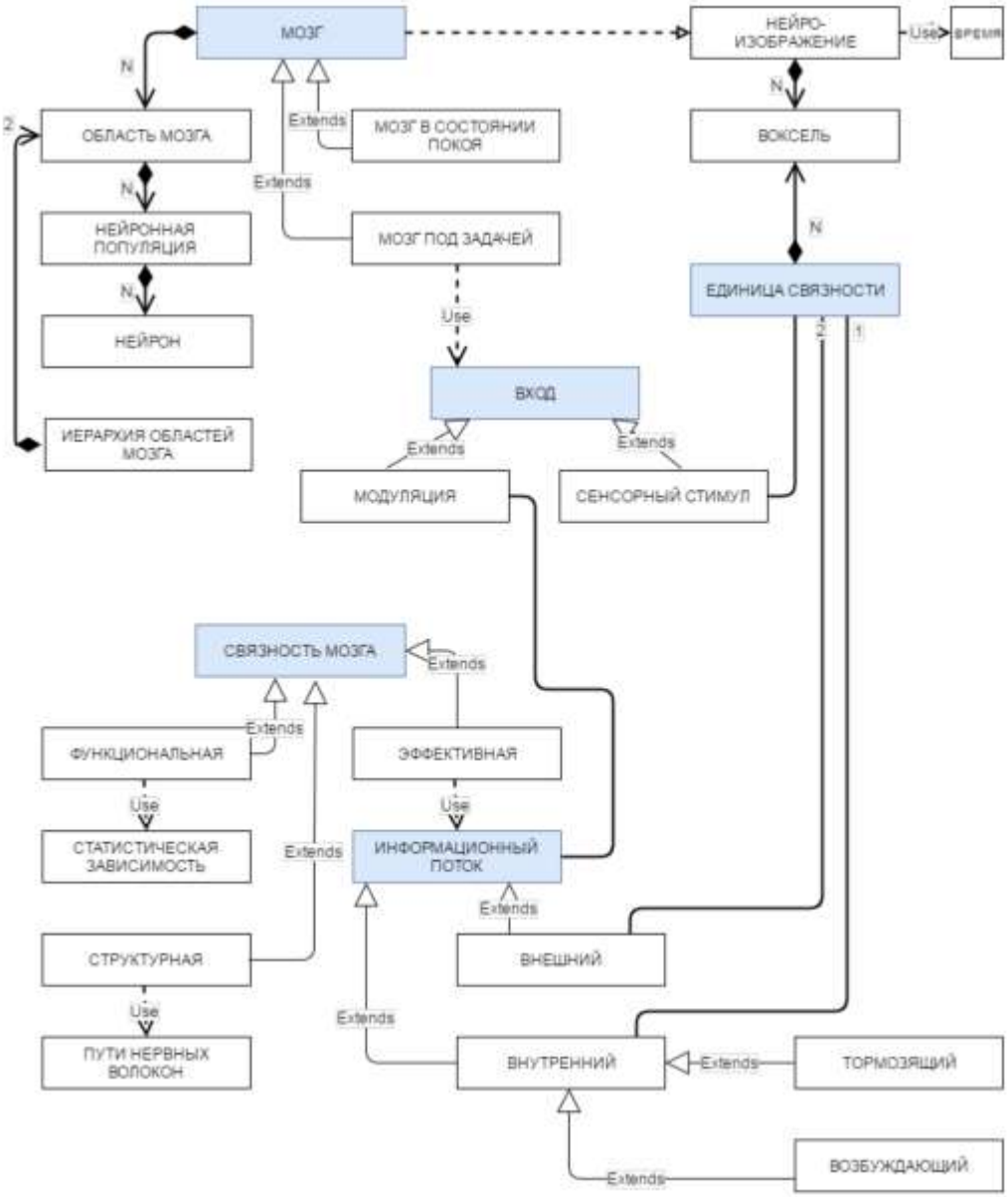


Рисунок 1

В каждом элементе массива (вокселе) содержится некоторое значение, отражающее силу сигнала в данном вокселе (чем больше значение, тем светлее пиксель на изображении). К определенному вокселю можно обратиться как к элементу массива по индексу.

Изображение, полученное методом фМРТ, состоит из последовательности трехмерных изображений и, таким образом, имеет четвертое измерение (временное) и содержит временные ряды для каждого вокселя.

Основным форматом данных в нейровизуализации методом фМРТ является формат NIFTI, содержащий в себе сами данные, а также информацию о том, как данные собирались со сканера.

Одной из главных причин сложности нервной системы является её сложная морфология, особенно взаимосвязность обрабатывающих нейронных элементов. Паттерны нейронной связности играют важную роль в определении функциональных свойств нейронов и нейронных систем. В более развитых нервных системах связность мозга можно рассматривать на разных масштабных уровнях. На микроуровне отдельные нейроны связываются друг с другом при помощи синаптических связей, на мезоуровне нейронные популяции объединяются в сети, а на макроуровне области мозга связываются нейронными проводящими путями, которые часто представляют собой пучки аксонов и служат для соединения относительно удаленных участков мозга или нервной системы.

Выделяются три вида связности мозга – структурная (или анатомическая связность), функциональная связность и эффективная связность.

Структурная связность описывает анатомические связи между нейронами посредством аксонов. Структурные связи специфичны и непостоянны. Структурная связность различна не только для

разных особей одного и того же вида, но и изменяется со временем у одного индивидуума. Она является относительно стабильной на коротких временных интервалах.

Согласованность работы отдельных областей оценить труднее, чем локализовать их активность. Один из возможных подходов – оценить функциональные связи, то есть статистические зависимости между активностью в различных областях мозга. Функциональная связность часто вычисляется на всех элементах системы, независимо от того, соединены ли данные элементы напрямую структурными связями. Функциональная связность также очень зависима от времени (может изменяться за десятые и сотые доли секунды). Также функциональная связность не связана очевидным образом с внутренней структурой модели или с влиянием одних областей мозга на другие.

Лучше всего интеграция оценивается в терминах эффективных связей, которые показывают, каким образом одна нейрональная система влияет (оказывает эффект) на другую. В некотором смысле эффективные связи – отражение истинных нейрональных процессов, в то время как функциональные связи – лишь статистически значимое сходство активности в различных областях мозга. Применительно к взаимодействиям между нейрональными популяциями эффективные связи – это каузальное взаимодействие между возбуждающими популяциями (являющимися причиной возбуждения или источником информации) и зависимыми от них или возбуждаемыми популяциями [8].

2.2 Пример множества априорных гипотез

Гипотезы являются ключевым элементом эксперимента. Формализация априорных гипотез может быть полезна, если потом потребуется заменить одну из них конкурирующей.

Исследование эффективной связности на задаче восприятия одушевленного движения опирается на базовые априорные гипотезы, которые в последующем сокращают пространство соревнующихся моделей:

H1: Область V5 – область зрительной системы, которая по предположению играет важную роль в восприятии любого типа движения;

H2: Область pSTS – область, которая по предположению больше активируется при восприятии одушевленного движения;

H3: По теории предсказательного кодирования по прямым связям передается ошибка предсказания, а по обратным – предсказание. Иерархически высший регион стремится уменьшить ошибку предсказания;

H4: Одушевленное движение менее предсказуемое, чем механическое;

H5: Внутренняя структура моделей не различается среди участников эксперимента;

H6: Есть разница в связности мозга (количестве информации, передаваемой между регионами) при восприятии одушевленного и неодушевленного движения;

H7: Входные стимулы по-разному влияют на исследуемые регионы.

2.3 Анализ методов эффективной связности фМРТ человека

Существует два разных подхода к поиску каузальных взаимодействий: моделирование и статистические оценки асимметричных метрик связанности двух и более процессов (временных рядов).

Известны следующие метрики для линейного случая: кросс-корреляция (англ. cross-correlation), условная взаимная информация (англ. conditional mutual information), трансферная энтропия (англ. transfer entropy, TE [14]), моментальный перенос информации (англ. momentary information transfer) и др.

Перечисленные метрики асимметричны при перестановке процессов местами и при обращении времени. Для их использования, как правило, не требуется никакого знания о системе и составляющих ее подсистемах, а также нет необходимости в априорных гипотезах и характере каузальных взаимодействий. Метрики теории информации могут быть применены к данным любой природы: от анализа финансовых рынков до геофизики и нейронаук. Отсюда вытекают основные преимущества и недостатки подобных метрик: универсальность; возможность первоначального анализа сложных систем, о которых ничего не известно; способны показать лишь наличие или отсутствие влияния одних процессов на другие, не предполагая под временными рядами никакого внутреннего процесса.

Моделирование представлено большим количеством методов, таких, как причинность по Грейнджеру (англ. Granger Causality, GC [15]), моделирование структурными уравнениями (англ. Structural Equation Modelling, SEM [16, 17]), динамическое моделирование причинности (динамическое каузальное моделирование, англ. Dynamic Causal Modelling, DCM [18]) и др.

В нейронауках наблюдаемые экспериментальные данные неинвазивных методов (электрическая активность на скальпе в ЭЭГ или изменение концентрации дезоксигемоглобина в фМРТ) являются лишь коррелятами истинной, интересующей экспериментатора активности нейронов, поэтому моделирование может быть очень полезным для понимания скрытых состояний и процессов.

Недостатки моделирования: зависимость от априорных знаний и предположений; методы плохо подходят для первоначального анализа больших массивов экспериментальных данных, так как нуждаются в проверяемых гипотезах.

3 Формализация потока работ

Схема потока работ представлена на Рис. 2.

Рисунок 2

Формирование и вывод априорных гипотез, цели и задачи исследования, предположение о методе, который будет использован для анализа. На этом этапе исследователи формируют задачу исследования, обозначают априорные гипотезы, если это необходимо, формируют представления о том, какими методами будет осуществляться анализ, какие инструменты будут использоваться и т. д.

Проектирование эксперимента. На данном этапе исследователи обдумывают и подготавливают эксперимент, опирающийся на априорные гипотезы и цели исследования. Например, эксперименты для проекта HCP спроектированы с помощью системы E-Prime.

Сбор данных. На данном этапе проводится эксперимент, участники эксперимента выполняют задания, в то время как исследователи собирают нейроданные с МРТ сканера.

Предобработка данных. На данном этапе данные, полученные на предыдущем этапе, подвергаются низкоуровневой предобработке, включающей в себя коррекцию движения, удаление пространственных артефактов и искажений, нелинейную регистрацию вокселей в пространстве MNI152, выделение масок и т. д.

Ядерное сглаживание. Гауссовское ядерное сглаживание – техника сглаживания изображений, наиболее часто используемая в нейровизуализации.

ФМРТ данные содержат много шума, но исследования показывают, что в основном этот шум гауссовский, т. е. случайный, независимый от вокеля к вокселю и ориентированный вокруг нуля. Таким образом, если усреднить интенсивности соседних вокселей, то шум будет стремиться к нулю, а сигнал – к некоторому среднему значению, отличному от нуля.

Таким образом, применяя пространственное сглаживание некоторой ядерной функцией, исследователи стремятся «размыть» изображение, смягчив жесткие края, снизив общую пространственную частоту, и улучшить отношение сигнал/шум:

$$Y(t) = \int K(t,s)X(s)ds,$$

где K – ядро интеграла, X – входной сигнал, Y – выходной сигнал.

Создание контрастов и регрессоров, подготовка дизайн-матрицы. Различные программы имеют разные методы настройки дизайн-матриц эксперимента, но все они основываются на том, что исследователь описывает набор различных экспериментальных условий (или эффектов) и указывает для каждого из них время начала и время окончания (либо время начала и продолжительность). Эксперименты могут иметь разные формы, от простейших блочных конструкций до сложных случайных событий со множеством условий. Основная гипотеза заключается в том, что некоторые воксели в мозге имеют статистически значимую активацию при некоторой комбинации экспериментальных условий и параметров.

Дизайн-матрица представляет собой матрицу, в которой строки представляют собой временные точки, а столбцы – смоделированные экспериментальные эффекты. Обычно исследователь модифицирует дизайн-матрицу, чтобы сделать модель более точно характеризующей активность мозга. Часто в матрицу добавляется константный регрессор для учета среднего значения сессии. Иногда к матрице добавляются линейные или полиномиальные дрейфы.

Иногда столбцы матрицы свертываются с некоторой функцией гемодинамического ответа для отражения размытости в сигнале [9].

Первый уровень анализа (по каждому субъекту). Большинство стандартных способов анализа фМРТ предполагает использование общей линейной модели (англ. General Linear Model, GLM), которая

по сути представляет собой расширение линейной множественной регрессии для случая одной зависимой переменной.

В ОЛМ исследователь задает предполагаемую модель ответов мозга при изменении некоторых экспериментальных условий, а затем проверяет истинность этой гипотезы. (Альтернативой ОЛМ может считаться метод главных компонент (англ. PCA)). Оценка общей линейной модели является основным статистическим этапом процесса анализа нейроданных, в котором исследуется, насколько заданная модель правильно интерпретирует реальные данные для каждого вокселя.

Стандартное уравнение GLM:

$$Y(t) = bX + \varepsilon,$$

где Y – изменяющиеся во времени интенсивности от одного вокселя, X – дизайн-матрица, ε – вектор нормально распределенных ошибок, а b – «бета-веса» – искомый параметр, вектор значений, показывающий, насколько значительным был вклад каждого эффекта экспериментального условия в объяснение значений в данном вокселе. Уравнение решается методом наименьших квадратов (англ. Ordinary Least Squares, OLS):

$$b = (X^T X)^{-1} X^T Y.$$

Если b-веса регрессора А значительно выше, чем b-веса регрессора В в данном вокселе, подтверждается гипотеза о том, что А обладает большим эффектом, чем В в этом вокселе.

После получения оценок эффектов и ошибок для каждого вокселя можно вычислить статистическую значимость для эффектов с использованием контрастов, определенных на предыдущем шаге, и получить контрастное изображение (англ. contrast image, beta image), содержащее информацию о величине интересующего эффекта: $t = \frac{c^T b}{Std(c^T b)}$.

При задании контрастов, призванных выделить наиболее активные воксели при определенном условии, тестируют гипотезу: H0: cb=0, где с – веса для b.

Второй уровень анализа (групповой). Главный вопрос группового анализа заключается в том, повторяется ли паттерн активности вокселей среди участников эксперимента. Главная цель состоит в том, чтобы найти групповую маску, которая бы выделяла воксели, активность которых была статистически значима среди всех участников для некоторого экспериментального условия.

На вход ОЛМ второго уровня, в отличие от ОЛМ первого уровня, в качестве вектора Y подаются контрастные изображения, полученные на предыдущем шаге. Дизайн-матрица на этом уровне анализа общая (групповая).

Есть несколько вариантов статистических тестов для группового анализа в зависимости от того, учитывается ли различие между субъектами (Random Effects vs Fixed Effects); относятся ли все субъекты к одной группе либо их нужно анализировать, разделив перед этим на несколько групп, например, на мужчин и женщин (one-sample t-test, two-sample t-test); какие и какое количество контрастов будет использоваться (t-test, F-test, ANOVA).

Выбор вокселей областей интереса, показавших наибольшую активацию, извлечение временных рядов для этих областей (локализация нейронной активности). На данном шаге воксели сопоставляются с атласом регионов мозга. Если гипотеза касалась определенных регионов мозга, то определяются координаты областей, интересующих исследователей. По статистическим картам предыдущего этапа выделяются пиковые воксели, находящиеся ближе всего к интересующим областям, после чего единицей связности (областью интереса) помечаются либо пиковый воксель, либо еще несколько окружающих его вокселей. Далее извлекаются временные ряды для этих областей интереса и усредняются, если вокселей в единице связности несколько.

Спецификация моделей. Спецификация моделей зависит от выбранных априорных гипотез и метода анализа. Каждый метод анализа сокращает пространство моделей [10].

Моделирование В DCM разделяется на два уровня – нейронный и BOLD. МРТ измеряет активность исследуемых областей не напрямую (нейронный уровень), а посредством измерения гемодинамических реакций, являющихся следствием нейронной активации (BOLD уровень), и основной идеей метода DCM является оценка параметров нейронного уровня с использованием априорно известной биофизической информации так, чтобы модель нейронной системы наилучшим образом предсказывала наблюдаемый BOLD-сигнал.

Далее под системой подразумевается набор элементов (нейронных популяций, областей), взаимодействующих друг с другом во времени и в пространстве.

DCM моделирует для каждой исследуемой популяции изменение абстрактного нейронного состояния (англ. state) во времени. Важны не сами значения этих состояний, а их динамика во времени.

Уравнение состояний для билинейной модели имеет вид:

$$\dot{z} = (A + \sum_{j=1}^{m} u_j B^j) z + Cu.$$

Таким образом, в DCM нейронная динамика моделируемой системы зависит от 4х параметров: A – внутренняя связность элементов системы (влияние нейронных популяций друг на друга в терминах причинности), B – контекстуальные входы (изменения силы связности), C – управляемые входы (сенсорные стимулы, источники активации модели), σ – задержка вызванной активности.

Моделируемая динамика нейронных популяций z (нейронный уровень) трансформируется в BOLD-сигнал y (BOLD уровень) с помощью гемодинамической модели, называемой моделью

«Воздушного шара». Комбинируя нейрональные z и гемодинамические y состояния в совместный вектор состояния x, а нейрональные и гемодинамические параметры – в совместный вектор параметров θ, получим полную прямую (порождающую) модель.

Важно отметить, что с помощью DCM можно получить два разных вывода [10].

Если интересны не конкретные параметры модели, но сама ее структура, то требуется заключение о пространстве моделей, например, если хочется узнать, имеет ли конкретная нейронная система последовательную или параллельную архитектуру, и при этом не требуется знать, касается ли контекстно-зависимая модуляция прямых или обратных связей или является ли механизм модуляции линейным или нелинейным. Или наоборот, если интересны нейрофизиологические механизмы, закодированные в конкретных параметрах данной модели, то требуется заключение о параметрах модели. Например, для некоторой модели хочется знать, будет ли конкретная связь оказывать с большей вероятностью возбуждающее или тормозящее воздействие на целевую область.

Перед проведением исследования методом DCM следует уточнить тип вывода, требуемый для данного вопроса. Этот выбор определяет последовательности шагов анализа данных [11].

Даже если исследователь заинтересован в выводе параметров модели, первым шагом обычно является Байесовский выбор модели (БВМ, англ. Bayesian Model Selection, BMS). BMS – это статистическая процедура по вычислению приближения доказательства модели (англ. model evidence).

Доказательство модели (или функцию маргинального правдоподобия) можно рассматривать как «святой грааль» сравнения моделей, которое количественно определяет свойства хорошей модели. В BMS модели обычно сравниваются по их байесовскому фактору, т. е. по соотношению их функций маргинального правдоподобия или, что эквивалентно, их разнице логарифмических доказательств.

Для вывода о пространстве модели BMS достаточно, но его можно применять по-разному. Можно, например, определить одну оптимальную модель либо разбить пространство моделей и сравнивать множества или семейства моделей, которые отличаются одним или несколькими структурными аспектами.

Выбор лучшей модели. Обычно на первом этапе специфицируются все вероятные модели, затем используется BMS для выбора оптимальной модели из всех альтернатив и, наконец, происходит оценка апостериорных и условных вероятностей параметров этой оптимальной модели. Для анализа данных одного субъекта вывод о каком-либо определенном параметре (или о линейных комбинациях параметров) довольно прост: нужно оценить апостериорную плотность интересующего параметра и количественно оценить вероятность того, что значение параметра больше или меньше некоторого порога. Для групповых исследований существует два варианта в зависимости от того, предполагается ли, что интересующие параметры имеют фиксированное распределение среди всех участников эксперимента (фиксированные эффекты, англ. FFX, Fixed Effects) или сами вероятностно распределены в популяции (случайные эффекты, RFX, Random Effects) (используется для поиска патологий и заболеваний среди исследуемых).

Цель выбора модели состоит в том, чтобы определить оптимальную модель из набора правдоподобных альтернатив, которая является наиболее полезной, то есть представляет наилучший баланс между точностью и сложностью и, таким образом, обеспечивает максимальную обобщаемость. Чтобы не потеряться в пространстве бесконечных возможных моделей, требуется тщательно определять размер пространства моделей.

При FFX предположении полезной метрикой является групповой Байесовский фактор (англ. Group Bayes Factor, GBF), который выражает функцию маргинального правдоподобия одной модели относительно функции маргинального правдоподобия другой модели, рассматривая группу в целом. У GBF есть простое определение: поскольку коэффициенты Байеса являются вероятностями, которые независимы между субъектами, GBF является продуктом отдельных факторов Байеса. При сравнении более двух моделей проще для каждой модели определить групповую логарифмическую функцию маргинального правдоподобия, что является суммой логарифмических доказательств по всем субъектам. На практике самым простым и информативным способом отображения группового логарифмического доказательства является построение гистограммы доказательств для каждой модели после вычитания лог-доказательства для модели с наименьшими доказательствами.

Нужно помнить, что доказательство модели определено в отношении одного конкретного набора данных. Это означает, что BMS нельзя применять к моделям, которые относятся к разным наборам данным. Важно иметь в виду, что любой результат, полученный BMS или любой другой процедурой выбора модели, выражает относительное утверждение о добротности модели, которое обусловлено рассмотренным пространством моделей [12].

Оценка параметров лучшей модели. Для вывода о параметрах модели необходимо оценить их апостериорные плотности распределения. Однако эти апостериорные оценки обусловлены выбранной моделью. По этой причине BMS обычно является обязательным этапом, даже если гипотеза касается значений параметров модели, а не структуры модели как таковой.

Для анализа групповых оценок параметров при FFX предположении обычно используют метод усреднения байесовских параметров (англ. Bayesian parameter averaging, BPA). Он эффективно вычисляет

совместную апостериорную плотность для всей группы, рассматривая апостериорную плотность одного субъекта как априорную для следующего. Метод производит единственную апостериорную плотность для всей группы, которая может быть использована для байесовского вывода. Альтернативами BPA считаются одномерный вариант BPA и простое усреднение временных рядов испытуемых в качестве этапа предварительной обработки (что возможно только в том случае, если стимулы появляются в одно и то же время для всех субъектов) [13].

Альтернативным подходом для вывода о параметрах модели является байесовское усреднение модели (англ. Bayesian model averaging, BMA). Этот подход отказывается от зависимости вывода о параметрах от выбранной модели. Вместо этого он использует все рассматриваемое пространство моделей и вычисляет средневзвешенные значения каждого параметра моделей, где взвешивание задается апостериорной вероятностью для каждой модели. Он представляет собой полезную альтернативу, когда ни одна из моделей не считается явно превосходящей все остальные. BMA также используется для сравнения оценок параметров между группами в тех случаях, когда BMS указала групповую разницу в отношении оптимальной модели.

4 Заключение

Работа посвящена исследованию методов проведения виртуального эксперимента. В качестве примера использована задача поиска эффективной связности фМРТ действия. В процессе исследования построена онтология предметной области, описаны характерные особенности данных, явно специфицированы используемые априорные гипотезы, формально описан поток работ поиска эффективной связности фМРТ человека.

Поддержка

Работа выполнена при поддержке РФФИ (грант 16-07-01028).

Литература

[1] Hillebrandt, H., Friston, K.J. et al.: Effective Connectivity During Animacy Perception – Dynamic Causal Modelling of Human Connectome Project Data. Scientific Reports, 9 (4:6240) (2014).

[2] Penny, W., Friston, K.J.: Statistical Parametric Mapping: The Analysis of Functional Brain Images. Academic Press, 656 p. (2006)

[3] Born, R.T., Bradley, D.C.: Structure and Function of Visual Aarea Mt. Annual Reviews Neuroscience, 28, pp. 157-189 (2005)

[4] Castelli, F., Happe, F.et al.: Movement and Mind: A Functional Imaging Study of Perception and Interpretation of Complex Intentional Movement Patterns. NeuroImage, 12, pp. 314-325 (2000)

[5] Van Essen, D. et al. The Human Connectome Project: A Data Acquisition Perspective. NeuroImage, 62, pp. 2222-2231 (2012)

[6] Reference Manual – Q2 Data Release (June 2013) | WU-Minn Consortium of the NIH Human Connectome Project http://www.humanconnectome. org/ documentation/Q2/Q2_Release_ Reference_Manual.pdf

[7] Glasser, M.F. et al.: The Minimal Preprocessing Pipelines for the Human Connectome Project. NeuroImage, 80, pp. 105-124 (2013).

[8] Van Essen, D., Ugurbil, K.: The Future of the Human Connectome [Review]. NeuroImage, 62 (2), pp. 1299-1310 (2012).

[9] http://mindhive.mit.edu/node/46

[10] Stephan, K.E. et al.: Ten Simple Rules for Dynamic Causal Modeling. NeuroImage, 49, pp. 3099-3109 (2010)

[11] Kasess, C.H. et al.: Multi-Subject Analyses with Dynamic Causal Modeling. NeuroImage, 49, pp. 3065-3074 (2010)

[12] Stephan, K.E., Penny, W.D. et al.: Bayesian Model Selection for Group Studies. NeuroImage, 46, pp. 1004-1017 (2010)

[13] Rosa, M. J., Friston, K., Penny, W.: Post-hoc Selection of Dynamic Causal Models. J. of Neuroscience Methods, 208, pp. 66-78 (2012)

[14] Schreiber, T. Measuring Information Transfer. Physical Review Letters, 85, p. 461 (2000)

[15] Friston, K. Dynamic Causal Modeling and Granger Causality Comments on: The Identification of Interacting Networks. The Brain Using fMRI: Model Selection, Causality and Deconvolution. Neuroimage, 58, pp. 303-305 (2011)

[16] Zhuang, J., LaConte, S. et al.: Connectivity Exploration with Structural Equation Modeling: an fMRI Study of Bimanual Motor Coordination. NeuroImage, 25, pp. 462-470 (2005)

[17] Pearl, J.: The Causal Foundations of Structural Equation Modeling. Chapter for R.H. Hoyle (Ed.), Handbook of Structural Equation Modeling, Guilford Press, 5, pp. 68-91 (2012)

[18] Friston, K.J., Harrison, L., Penny, W.: Dynamic Causal Modeling. Neuroimage, 19 (4), pp. 1273-1302 (2003)

Специфические методы анализа данных

Specific data analysis techniques

Формирование исторической справки по корпусу новостей с учетом структуры динамики развития новостного сюжета

© М.М. Тихомиров © Б.В. Добров

Московский государственный университет имени М.В. Ломоносова,
Москва, Россия

tikhomirov.mm@gmail.com dobrov_bv@srcc.msu.ru

Аннотация. Описаны проведенные исследования по тематике формирования исторической справки. Разработаны алгоритмы и реализована программная система, позволяющая автоматически создавать историческую справку по корпусу новостных статей для выбранного новостного документа. Проведено исследование трех новых факторов, учитывающих структуру динамики развития новостного сюжета.

Ключевые слова: обзорное реферирование, историческая справка, информационный поиск.

Using News Corpora for Temporal Summary Formation

© Mikhail Tikhomirov © Boris Dobrov

Lomonosov Moscow State University,
Moscow, Russia

tikhomirov.mm@gmail.com dobrov_bv@srcc.msu.ru

Abstract. The paper describes the research carried out on the subject of the formation of the temporal summary. Algorithms have been developed and a software system has been implemented that allows you to automatically create a timeline summary for the body of news articles for the selected news document. A study of three new factors, taking into account the structure of the dynamics of news story development.

Keywords: timeline summarization, multi-document summarization, information retrieval.

1 Введение

В связи с взрывным ростом количества информации в интернете возникает задача выделения и автоматического обобщения полезной информации в поступающем потоке данных.

Востребованными задачами являются задачи реферирования новостных сюжетов – множества новостных сообщений различных источников, посвященных описанию некоторого события. Такие задачи часто решаются новостными агрегаторами, например, Яндекс.Новости [17], для более полного представления описания произошедшего события. Типичное «время жизни» новостного сюжета (время активного обсуждения произошедшего события) обычно сутки–двое.

Отметим, что некоторые новостные сюжеты имеют «историю» в виде множества предшествующих событий, произошедших в различные моменты времени и в той или иной мере связанных между собой.

Для таких длительных сюжетов, где сама их длительность (повторное возвращение к одной и той же теме) в определенной мере свидетельствует об их значимости, является актуальной задача формирования «исторических справок».

Историческая справка – это тип обзорного реферата (обзорной аннотации), включающего последовательное изложение существенных деталей исследуемого сюжета. Подобная аннотация может содержать в себе основные этапы, события и факты исходного сюжета. Построение подобных аннотаций представляет собой сложную работу, которую выполняют журналисты или аналитики, и, соответственно, автоматизация подобного процесса является востребованной задачей.

В рамках данной работы рассмотрены проблемы и решения при автоматическом построении исторических справок.

Рассматривается ситуация, когда пользователя новостного агрегатора заинтересовала какая-то новость (новостное сообщение), и он хочет получить историческую справку по сюжету, обсуждаемому в данном новостном сообщении, т. е. результатом должен быть упорядоченный по времени перечень описаний произошедших ранее ключевых событий.

Задача рассматривается как задача обзорного реферирования (multi-document summarization) по запросу на представительной коллекции новостных

Труды XIX Международной конференции «Аналитика и управление данными в областях с интенсивным использованием данных» (DAMDID/ RCDL'2017), Москва, Россия, 10–13 октября 2017 года

документов. В качестве запроса рассматривается текст новостного сообщения.

На корпусе из 2 миллионов новостных статей на русском языке за первую половину 2015 года была разработана и реализована система, позволяющая автоматизировать процесс построения исторической справки. Проведено исследование трех новых факторов, позволяющих за счет учета структуры новостного корпуса улучшить результаты работы системы. Оценка производилась на 15 новостных сюжетах, из которых для 5 эталонные аннотации были сформированы одним из авторов, а другие 10 взяты с сайта interfax.ru

2 Обзор

2.1 Задача обзорного аннотирования

В настоящее время предложено достаточно большое количество методов автоматического обзорного реферирования [3]. Известны методы как с использованием больших лингвистических онтологий [15], в том числе автоматически пополняемых в процессе анализа [12], так и на основе статистических свойств текстов [16], машинного обучения [13, 17].

Существенными проблемами при составлении аннотации новостного кластера являются [3, 7, 11]:

- обеспечение полноты представления информации, в том числе наиболее свежей информации;
- снижение повторов при представлении информации;
- обеспечение связности и понятности представляемой информации.

Для определения избыточности в порождаемых аннотациях используются различные меры сходства между предложениями. Одним из распространенных подходов является предварительная кластеризация выделение близких по содержанию кластеров предложений [6]. Другим подходом для уменьшения избыточности являются сравнение предложений-кандидатов с предложениями, уже попавшими в аннотацию, и оценка новой (непохожей) информации, например, подход Maximal Marginal Relevance (MMR) [2].

2.2 Историческая справка

Задача построения исторических справок имеет ряд отличий от стандартной задачи обзорного реферирования.

Сначала необходимо определить документы, по которым будет строиться аннотация. Если стандартный новостной сюжет обычно образован близкими документами, посвященными одному событию, которые могут быть получены применением одного из известных методов кластеризации [10, 14].

Для больших коллекций применение методов кластеризации не оправдано. Во-первых, такую задачу придется решать многократно на огромных коллекциях документов. Во-вторых, степень близости между документами, которые описывают далекие по времени, но связанные события, может быть значительно меньше по стандартным мерам сходства.

Требуется выявлять наиболее характерные объекты [1, 9], например, учитывая структурные особенности потока документов [5, 8].

3 Постановка задачи

3.1 Общее описание

Задача построения исторической справки ориентирована на запрос. В самом общем случае пользователь в качестве запроса имеет новостной документ, поэтому данная задача будет рассматриваться как задача автоматического построения аннотации описанного типа по запросу в виде текстового документа. На выходе работы системы должна быть аннотация из n предложений. Связность между предложениями не требуется.

Как пример построенной исторической справки можно рассмотреть аннотацию (таблица 1), построенную по событию, связанному с крушением самолета в Тайване.

Таблица 1. Крушение самолета на Тайване

1	*Самолет ATR 72 авиакомпании TransAsia потерпел крушение 4 февраля на Тайване.*
2	*Операция по поиску жертв крушения самолёта TransAsia Airways завершена, в результате происшествия погибли 35 человек.*
3	*Члены экипажа самолета авиакомпании TransAsia Airways, потерпевшего крушение в феврале на Тайване, отключили работающий двигатель, после того, как второй перестал работать*
…	*…*
n	*Совет по авиационной безопасности Тайваня опубликовал отчет о крушении самолета компании TransAsia Airways в феврале этого года, в результате которого погибли 35 человек.*

В цели работы входит исследование влияния различных факторов на качество построения аннотации, поэтому необходим набор эталонных аннотаций, на которых будет оцениваться качество работы системы.

3.1 Математическая постановка задачи

Описанную выше задачу можно формализовать следующим способом: имеются набор запросов $Q = \{q_1, q_2, \dots, q_m\}$ и ассоциированный с ним набор эталонных аннотаций $D_g = \{D_g^{q_1}, D_g^{q_2}, \dots, D_g^{q_m}\}$. Система в ответ на запросы Q алгоритмом A генерирует набор исторических справок $D_A = \{D_A^{q_1}, D_A^{q_2}, \dots, D_A^{q_m}\}$.

Тогда задача построения исторической справки сводится к задаче максимизации функционала

$$\frac{\sum_{i=1}^{i=|Q|} M(D_A^{q_i}, D_g^{q_i})}{|Q|} \to \max, \quad (1)$$

где M – функция близости между аннотациями. Максимизация происходит по выбору алгоритма A и по всем параметрам выбранного алгоритма.

4 Предлагаемый подход

4.1 Исследуемые факторы

В рамках работы исследовались следующие факторы:

- стратегия расширения запроса;
- учет временного характера новостных сюжетов.
- учет структуры новостной статьи в виде перевернутой пирамиды.

4.2 Стратегия расширения запроса

Информации, которую можно получить из запроса-документа, может быть не достаточно, чтобы эффективно построить историческую справку. Этот факт является следствием того, что большинство новостных статей является не общим описанием события, а обсуждением какого-то частого происшествия или факта. Чтобы избежать подобной проблемы, был разработан алгоритм, использующий кластер близких запросу документов. Алгоритм:

1. Для запроса-документа на основе статистической информации по коллекции (индекс) строится вектор наиболее весомых по tf-idf лемм (нормализованных словоформ) документа.
2. По построенному вектору происходит поиск близких документов в коллекции.
3. По кластеру извлеченных документов происходит анализ важности лемм на основе tf-idf:
    a. Для каждого документа рассматриваются лучшие t лемм.
    b. Происходит ранжирование лемм на основании частоты встречаемости в лучших *t* леммах каждого документа.
    c. Из сортированного списка выбирается *k* наиболее весомых лемм.
4. Повторяются пункты 2–3 (повторное расширение запроса).
5. На выходе имеется вектор из *k* лемм, который отражает семантику документа-запроса.

Как пример работы модуля расширения запроса можно рассмотреть этапы работы алгоритма на новостной статье, посвященной теракту в Париже (порядок в списке обратный по отношению к весу слова):

Олланд назвал нападение на Charlie Herbo терактом
Президент Франции Франсуа Олланд назвал терактом нападение на сотрудников сатирического журнала Charlie Herbo в центре Парижа. По последним данным, в результате стрельбы погибли 11 человек, еще четверо находятся в критическом состоянии. …

Первичный запрос, полученный на этапе 1:

1. *Posten, Jyllands-posten, Jyllands, Herbo, Charlie, Олланд.*

Единожды расширенный запрос, после этапа 3:

2. *Перепечатать, Скандальная, Еженедельник, Карикатура, Олланд, Сатирический, Теракт, Charlie, Herbo.*

Дважды расширенный запрос после этапа 5:

3. *Журнал, Мухаммед, Сатирический, Атака, Пророк, Теракт, Париж, Карикатура, Олланд, Herbo, Charlie.*

Как видно, последний вариант включает в себя наиболее важные элементы.

4.3 Учет структуры новостной статьи в виде перевернутой пирамиды.

Рисунок 1 Перевернутая пирамида «идеального» новостного сообщения

Стратегия написания качественной новостной статьи часто опирается на структуру вида «перевернутая пирамида», Рис. 1.

В дополнительной информации часто встречается описание произошедших ранее событий по теме документа.

Учет данной структуры происходит в 2 аспектах:

1. Построение графа из документов, близких к запросу, где ребром является неявная ссылка между окончанием одной статьи и началом другой статьи, которая была опубликована ранее.
2. Повышение веса предложений, которые располагаются в верхней части новостной статьи и нижней части. Выделение нижней части происходит из-за того, что предложения оттуда часто резюмируют информацию из заголовков других статей.

Алгоритм работы первого способа учета структуры «перевернутая пирамида» выглядит следующим образом:

1. Для набора документов D происходит построение матрицы близости между окончаниями и началами документов.
2. При превышении заданного порога считается, что присутствует ссылка между документами D_i и D_j.
3. На построенном графе происходит ранжирование документов путем использования известного алгоритма LexRank [4]. Веса документов нормируются.

4. Для наиболее весомых документов производится описанная ранее операция построения расширенного запроса.

5. Итого, на выходе имеется ранжированный список документов D и набор из p новых запросов, учет которых будет осуществлен совместно с учетом временной структуры новостного сюжета.

Второй способ учета структуры перевернутой пирамиды реализован в функции ранжирования итоговых предложений, раздел 4.6.

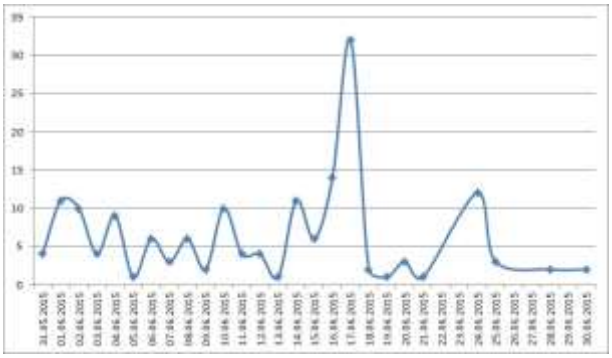


Рисунок 2 Зависимость количества публикаций новостного сюжета от времени

4.4 Учет временного характера новостных сюжетов

Так как любое событие зависит от времени, то публикации и количество публикаций тоже зависят от времени. Как пример, на Рис. 2 изображен график зависимости публикаций по событию «Землетрясение в Непале». Чтобы учесть данный фактор, для набора документов D происходит следующее:

1. Вся временная шкала события разбивается по суткам с метками $T = \{t_1, t_2, \dots, t_n\}$.
2. На основании информации о дате публикации документа каждый документ получает метку из T.
3. Происходит фильтрация дней с малым количеством публикаций. Это происходит за счет анализа количества публикаций для метки t_i к максимальному количеству публикаций в любой день и суммарному количеству публикаций.
4. На выходе имеется сортированный список, где каждый элемент имеет метку t_i из T и набор документов $D_i \in D$.

Помимо прочего, происходит отображение всех ранее построенных расширенных запросов на метки t_i из T, Рис. 3.

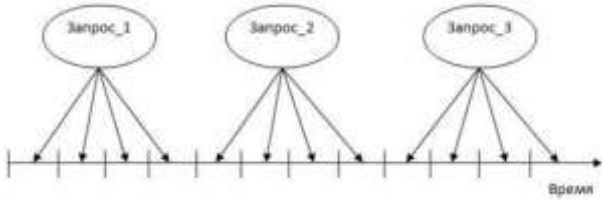


Рисунок 3 Отображение запросов на шкалу времени

4.5 Схема работы программой системы

Описанные в пунктах 4.1 факторы реализуются на различных этапах работы системы. Общая схема работы представлена на Рис. 4.

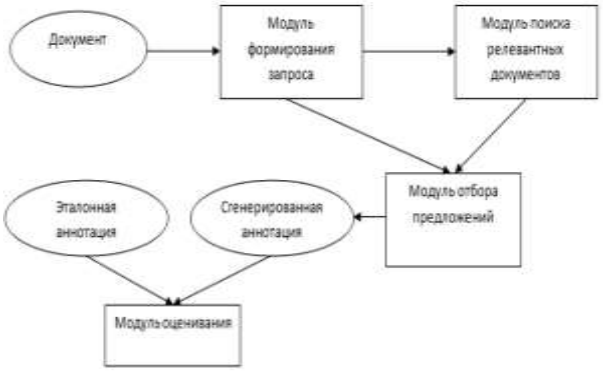


Рисунок 4 Схема работы системы

4.6 Модуль поиска релевантных документов

Поиск релевантных документов происходит путем поиска близких документов для построенного запроса на этапе формирования запроса, описанным в пункте 4.2. Использовалась поисковой машины NearIdx 8, разработанная ООО «Лаборатория информационных исследований».

4.7 Модуль отбора предложений

Данный модуль занимается непосредственно ранжированием предложений из извлеченных документов.

Ранжирование происходит модифицированной версией алгоритма MMR, которая прямо или косвенно учитывает все факторы, описанные в 4.1:

$$MMRT_{s_i^t} = INC_{s_i^t} - DEC_{s_i^t}, \quad (2)$$

где $INC_{s_i^t}$ – член, описывающий положительную составляющую формулы, которая зависит от близости предложения к запросу, веса документа, из которого взято предложение, и позиции предложения в документе;

$$INC_{s_i^t} = (1 + \alpha * I_i) * \gamma * \lambda * Sim(Q^t, S_i^t), \quad (3)$$

$$\gamma = 1 - 0.5 * \sin(\frac{i * \pi}{|D_s|}). \quad (4)$$

Параметры α и λ являются настраиваемыми параметрами алгоритма, I_i – вес документа D_s, в который входит предложение под индексом i, S_i^t – оцениваемое предложение под индексом i и с временной меткой t, Q^t – запрос, отображенный на эту временную метку, γ – слагаемое, понижающее вес предложений из середины документа.

Слагаемое $DEC_{s_i^t}$ – штрафное. Оно зависит от близости к уже извлеченным предложениям:

$$DEC_{s_i^t} = (1 - \lambda) * \max_{S_j \in S} Sim(S_j, S_i^t), \quad (5)$$

S_j – одно из извлеченных предложений, S – множество всех уже извлеченных предложений.

Обработка множества предложений, пришедших из модуля поиска релевантных документов, происходит в хронологическом порядке, на каждом этапе обрабатывается подмножество $D_i \in D$, связанное с меткой $t_i \in T$. Для каждого этапа имеется ограничение на извлечение максимум K предложений за сутки.

4.8 Мера близости

На различных этапах работы программной системы есть ряд моментов, когда вычисляется мера близости между предложениями. В работе использовались два подхода к расчету близости, использующих косинусную меру близости:

$$Sim_{cos}(S_i, S_j) = \frac{(S_i,\ S_j)}{|S_i|*|S_j|}. \quad (6)$$

Для расчета близости на этапе ранжирования предложений для них использовалось стандартное векторное представление, полученное из индекса, где вес элемента – это tf-idf.

Для расчета близости между окончаниями и началами новостных статей (на этапе построения графа) использовались вектора, полученные с помощью word2vec модели, обученной на всей коллекции документов.

5 Оценивание

5.1 Метрики оценивания

Оценивание работы системы происходило на нескольких метриках: ROUGE-1 и ROUGE-2, полноте по предложениям (8) и комбинированной метрики (9):

$$ROUGE - N = \frac{|N_A \cap N_g|}{|N_g|}, \quad (7)$$

где N_A – множество n-грамм словоформ для построенных аннотаций, N_g – для эталонных аннотаций;

$$P^{sent} = \frac{|S_A \equiv S_g|}{|S_g|}, \quad (8)$$

где S_A – множество предложений из построенных аннотаций, S_g – из эталонных аннотаций, а $\equiv$ понимается в том смысле, что в результирующем $S_A \equiv S_g$ остаются только те предложения из S_A, эквивалент которых есть в $S_{g;}$.

$$V^{comb} = 0.8 * R1 + R2 + 2 * P^{sent}, \quad (9)$$

где RN – сумма ROUGE-N и ее F-мера аналога ROUGE-NF.

5.2 Подготовка данных для процедуры оценивания

Так как для процедуры оценки качества работы системы необходим тестовый набор аннотаций, в рамках исследования были вручную подготовлены исторические справки. Процедура формирования такой коллекции происходила следующим образом:

1. На первом этапе происходил отбор ярких событий, которые активно освещались в прессе за период начала 2015 года.
2. Далее для большинства событий на информационном ресурсе interfax осуществлялся поиск соответствующего сюжета. Пример – на Рис. 5.
3. Если соответствующего сюжета на interfax нет, происходили изучение материалов по теме и формирование исторической справки на основе прочитанных документов.
4. Сюжеты просматривались в хронологическом порядке и производился отбор наиболее информативных предложений.
5. На основе отобранных предложений составлялись исторические справки, размер которых, в среднем, около 15 предложений.

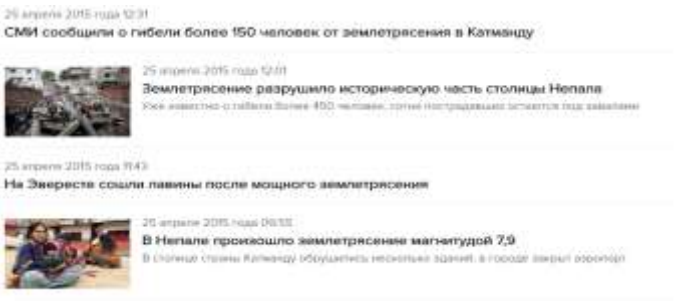


Рисунок 5 Отрывок сюжета с interfax (http://www.interfax.ru/story/151/page_3)

Итого, в результате построенная тестовая коллекция содержит в себе исторические справки по 15 событиям.

6 Результаты

Оценивались 6 конфигураций системы:

1. baseline – простой подход к аннотированию, без учета рассмотренных факторов, с использованием в качестве метода ранжирования обычного MMR;
2. querry-ex – добавление к baseline стратегии расширения запроса, но без повторного расширения запроса;
3. double-ex – querry-ex + двойное расширение запроса;
4. temporal – double-ex + учет временного характера сюжета;
5. importance – temporal + учет структуры перевернутой пирамиды;
6. full – importance + расчет близости на этапе построения графа происходит с помощью word2vec модели.

Каждая конфигурация настраивалась для получения максимального результата по всем внутренним параметрам системы (см. таблицу 2).

Результат измерений качества конфигураций можно увидеть в Таблице 3.

Таблица 2 Параметры системы

Название	Описание
SoftOr	Значение параметра soft_or_coef для поисковой машины.
KeepL	Количество лемм, выбираемых при построении первичного запроса.
KeepT	Количетсво терминов, выбираемых при построении первичного запроса.
DocCount	Значение параметра doccnt при построении расширенного запроса.
QuerrySize	Размер итогового расширенного запроса.
TopLemms	Количество наиболее значимых лемм, извлекаемых в работе

Название	Описание
	алгоритма построения расширенного запроса.
DocCount	Значение параметра doccnt при поиске релевантных документов.
MinSentSize	Минимальный размер предложения.
MaxSentSize	Максимальный размер предложения.
MinLinkScore	Минимальное значение близости окончания и заголовка документа для выявления ссылки.
Power MethodDFactor	Параметр D в алгоритме LexRank.
Power MethodEps	Параметр eps в алгоритме LexRank.
Lambda	Значение параметра λ для $MMRT$.
Alpha	Значение параметра α для $MMRT$.
MaxDaily AnswerSize	Максимальное количество предложений, извлекаемых за сутки.
Doc Boundary	Порог, позволяющий отобрать наиболее важные документы.
Init QuerrySize	Количество лемм, которые используются для повторного расширения запроса.

Таблица 3 Результаты оценивания конфигураций

Конфигурация	R1	R2	P^{sent}	V^{comb}
baseline	0.499	0.136	0.205	1.153
querry-ex	0.529	0.147	0.216	1.276
double-ex	**0.567**	**0.164**	0.260	**1.425**
temporal	0.564	0.162	0.251	1.400
importance	0.548	0.158	**0.261**	1.395
full	**0.566**	**0.162**	**0.262**	**1.433**

Полужирным шрифтом выделены по два лучших результата по каждой метрике.

Из Таблицы 3 можно сделать выводы, что наибольший вклад дало двойное расширение запроса. Факторы временной зависимости событий и структуры новостной статьи показывают неплохие результаты при совместном использовании. Также важную роль играет метрика близости, которая используется на каждом этапе решения.

Таблица 4 Отрывок исторической справки на тему падения самолета на Тайване

11.02.2015	Transasia Airways выплатит родственникам жертв авиакатастрофы на Тайване по 470 тыс.
11.02.2015	Трагедия на Тайване, одна пятая пилотов тайваньской авиакомпании Transasia не прошли тест на профпригодность.
12.02.2015	Спасатели завершили операцию по поиску жертв крушения самолёта авиакомпании Transasia Airways, который потерпел крушение 4 февраля на Тайване.
01.07.2015	Экипаж разбившегося на Тайване самолета Transasia Airways отключил двигатели после потери мощности.
02.07.2015	Самолет Transasia потерпел крушение 4 февраля на Тайване, потому что пилот по ошибке отключил работающий двигатель, когда второй двигатель заглох.

В качестве примера итоговой аннотации можно рассмотреть отрывок аннотации по упомянутому ранее событию падения самолета на Тайване в Таблице 4.

7 Заключение

Проведены исследования по тематике построения исторических справок. Были рассмотрено три фактора, которые могут влиять на качество построения аннотаций. Получены количественные и качественные результаты.

По результатам проведенных исследований оказалось, что выбор стратегии расширения запроса оказывает наибольшее влияние на качество построение аннотации подобного типа. Учет временного характера сюжета совместно с учетом структуры новостной статьи также улучшает результаты по метрикам P^{sent} и V^{comb}, что говорит о том, что данные факторы способны положительно влиять на качество построения исторических справок.

Литература

[1] Binh Tran, G., Alrifai, M., Quoc Nguyen, D.: Predicting Relevant News Events for Timeline Summaries. Proc. of the 22nd Int. Conf. on World Wide Web. ACM. pp. 91-92 (2013)

[2] Carbonell, J., Goldstein, J.: The Use of MMR, Diversity-based Reranking for Reordering Documents and Producing Summaries. Proc. of the 21st Annual Int. ACM SIGIR Conf. on Research and Development in Information Retrieval. ACM. pp. 335-336 (1998)

[3] Dang, H.T.: Overview of DUC 2006. Proc. of the document understanding Workshop. Presented at HLT-NAACL 2006 (2006). http://duc.nist.gov/pubs/2006papers/duc2006.pdf

[4] Erkan, G., Radev, D.R.: Lexrank: Graph-based Lexical Centrality as Salience in Text Summarization. J. of Artificial Intelligence Research, (22), pp. 457-479 (2004)

[5] Hu, P., Huang, M.L., Zhu, X.Y.: Exploring the Interactions of Storylines from Informative News Events. J. of Computer Science and Technology, 29 (3), pp. 502-518 (2014)

[6] Radev, D., Jing, H., Budzikowska, M.: Centroid-based Summarization of Multiple Documents: Sentence Extraction, Utility-Based Evaluation, and User Studies. Proc. of the 2000 NAACL-ANLP

Workshop on Automatic summarization. Seattle. pp. 21-30 (2000)

[7] Radev, D., McKeown, K., Hovy, E.: Introduction to the Special Issue on Summarization. Computational linguistics, 28 (4). pp. 399-408 (2002)

[8] Shahaf, D., Guestrin, C.: Connecting Two (or Less) dots: Discovering Structure in News Articles. ACM Transactions on Knowledge Discovery from Data (TKDD). 5 (4), pp. 24-54 (2012)

[9] Tran, G., Alrifai, M., Herder, E.: Timeline Summarization from Relevant Headlines. Hanbury A., Kazai G., Rauber A., Fuhr N. (eds) Advances in Information Retrieval. ECIR 2015. Lecture Notes in Computer Science, 9022. Springer, Cham. pp. 245-256 (2015). doi: 10.1007/978-3-319-16354-3_26

[10] Yan, R. et al.: Evolutionary Timeline Summarization: a Balanced Optimization Framework via Iterative Substitution. Proc. of the 34th Int. ACM SIGIR Conf. on Research and Development in Information Retrieval. Beijing, China. July 24– 28, 2011. ACM. pp. 745-754 (2011). doi: 10.1145/2009916.2010016

[11] Абрамова, Н.Н., Абрамов, В.Е.: Автоматическое составление обзорных рефератов новостных сюжетов. Труды 9-ой Всерос. науч. конф. «Электронные библиотеки: перспективные методы и технологии, электронные коллекции» – RCDL'2007, Переславль-Залесский, Россия. сс. 131-141 (2007)

[12] Алексеев, А.А., Лукашевич, Н.В.: Автоматическое порождение обновления к аннотации новостного кластера. Труды 12й Всерос. науч. конф. «Электронные библиотеки: перспективные методы и технологии, электронные коллекции» – RCDL'2010, Казань, Россия. сс. 81-91 (2010)

[13] Браславский П., Густелев, В.: Система автоматического реферирования новостных сообщений на основе машинного обучения. Труды Девятой Всерос. науч. конф. – RCDL'2007, Переславль-Залесский, Россия. Сс. 142-147 (2007)

[14] Добров, Б.В., Павлов, А.М.: Исследование качества базовых методов кластеризации новостного потока в суточном временном окне. Труды 12-й Всерос. науч. конф. «Электронные библиотеки: перспективные методы и технологии, электронные коллекции» – RCDL'2010, Казань, Россия. сс. 287-295 (2010)

[15] Лукашевич, Н.В., Добров, Б.В.: Автоматическое аннотирование новостных кластеров на основе тематического представления. Компьютерная лингвистика и интеллектуальные технологии: По материалам ежегодной Межд. конф. «Диалог 2009» (Бекасово, 27–31 мая 2009 г.). М.: РГГУ, Вып. 8 (15), сс. 299-305 (2009)

[16] Тарасов, С.Д.: Исследование и оптимизация параметров алгоритма Manifold Ranking на основе метрики автоматической оценки качества обзорного реферирования ROUGE-RUS. Труды XI Всерос. науч. конф. «Электронные библиотеки. Перспективные методы и технологии, электронные коллекции». Петрозаводск. сс. 86-93 (2009)

[17] Шаграев, А.: Автоматическое аннотирование новостного потока. Семинар: Natural Language Processing (автоматическая обработка естественного языка). Яндекс. 26.11.2011 (2011). https://www.slideshare.net/NataliaOstapuk/ ss-10380447?ref=http://nlpseminar.ru/lecture54/

Фрактальные методы в информационных технологиях обработки, анализа и классификации больших потоков астрономических данных

© А.В. Мышев © А.В. Дунин

Национальный исследовательский ядерный университет (МИФИ) – Обнинский институт атомной энергетики (ИАТЭ),
Обнинск, Россия

mishev@iate.obninsk.ru

Аннотация. Рассмотрена фрактальная парадигма построения моделей и логических схем алгоритмов и процедур информационных технологий обработки, анализа и классификации больших потоков астрономических данных об орбитах и траекториях малых тел. Методология построения таких моделей и схем основана на построении оценок критериев близости и связанности орбит и траекторий в пространстве возможных состояний с использованием соответствующего математического аппарата фрактальных размерностей. Логическая, алгоритмическая и содержательная сущности фрактальной парадигмы заключаются в следующем. Во-первых, обработка и анализ потока данных орбит и траекторий состоят в том, чтобы определить, образует ли он фрактальную структуру. Если да, то определить центры фрактальной связанности потока и получить оценки индекса информационной связанности орбит или траекторий. Во-вторых, нужно выделить монофрактальные структуры в потоке и классифицировать их по признаку принадлежности к классам перколирующего фрактала или фрактального агрегата.

Ключевые слова: связанность орбит, фрактальные меры, фрактальная размерность, перколирующий фрактал, фрактальный агрегат.

Fractal Methods in Information Technologies for Processing, Analyzing and Classifying Large Flows of Astronomical Data

© A.V. Myshev © A.V. Dynin

National Research Nuclear University MEPhI (IATE),
Obninsk, Russia

mishev@iate.obninsk.ru

Abstract. The fractal paradigm of constructing models and logical schemes of algorithms and procedures for information processing, analysis and classification of large flows of astronomical data on the orbits and trajectories of small bodies is considered. The methodology for constructing such models and schemes is based on the construction of estimates of proximity and connectivity criteria for orbits and trajectories in the space of possible states using the corresponding mathematical apparatus of fractal dimensions. The logical, algorithmic and substantial essence of the fractal paradigm is as follows. Firstly, the processing and analysis of the data flow of orbits and trajectories is to determine whether it forms a fractal structure. If yes, then it determines the centers of fractal connectivity of the flow and obtains estimates of the index of information connectivity of orbits or trajectories. Secondly, it is necessary to isolate the monofractal structures in the flow and classify them according to the attribute of belonging to the classes of a percolating fractal or a fractal aggregate.

Keywords: connectedness orbits, fractal measures, fractal dimension, percolating fractal, fractal aggregate.

1 Введение

Рассмотрены новые подходы к построению моделей и логических схем алгоритмов и процедур информационных технологий обработки, анализа и классификации больших потоков астрономических данных об орбитах и траекториях малых тел. Методология построения таких моделей и схем основана на построении оценок критериев близости и связанности орбит и траекторий в пространстве возможных состояний с помощью соответствующего математического аппарата фрактальных

Труды XIX Международной конференции «Аналитика и управление данными в областях с интенсивным использованием данных» (DAMDID/ RCDL'2017), Москва, Россия, 10–13 октября 2017 года

размерностей. Логическая, алгоритмическая и содержательная сущности методов и технологий фрактальной парадигмы заключаются, во-первых, в обработке и анализе потока данных орбит и траекторий с тем, чтобы определить, образует ли он фрактальную структуру (если да, то необходимо определить центры фрактальной связанности потока и получить оценки индекса информационной связанности орбит или траекторий), во-вторых, в выделении монофрактальных структур в потоке и классификации их по признаку принадлежности к классам перколирующего фрактала или фрактального агрегата. Фрактальная парадигма в методологии разработки и реализации информационных технологий обработки, анализа и классификации больших потоков данных, в отличие от традиционных методов и способов [1–4], позволяет учитывать как свойства регулярности и нерегулярности структуры пространства состояний информационной шкалы данных потока, так и их динамическую и информационную связанность.

Рассматриваемые логические схемы алгоритмов и процедур технологий обработки, анализа и классификации больших потоков данных построены на основе теории фрактальных размерностей пространственных и временных структур, алгоритмическая и содержательная сущность которых заключается в следующем. Во-первых, обработка потока данных состоит в том, чтобы определить, образует ли он фрактальную структуру. Если да, то необходимо определить центры фрактальной связанности потока данных и получить оценки индекса информационной связанности. Во-вторых, алгоритмы и процедуры технологий анализа и классификации обработанного потока позволяют выделить монофрактальные структуры, если поток образует мультифрактал, и классифицировать их по признаку принадлежности к классам перколирующего фрактала или фрактального агрегата, а также оценить меру расхождения между геометрическими и информационными фрактальными размерностями, как индикатора единства количественных и качественных характеристик потока.

2 Астрономические данные (обработка, анализ и классификация)

2.1 Постановка задачи

Для обработки потоков астрономических данных об орбитах и траекториях объектов космического пространства получены новые критерии их близости. В качестве критерия близости двух орбит введена количественная оценка фрактальной меры на множестве по-парных расстояний между соответствующими парами точек орбит. Такая оценка фрактальной меры является критерием связанности двух орбит, посредством которого отражается степень их геометрической и информационной близости. Основная посылка и смысл введенной сущности объясняются и поясняются следующей логической схемой и алгоритмом.

Во-первых, на орбитах определяются реперные (или опорные) точки, в качестве которых могут выступать перигелий или афелий орбиты. Во-вторых, относительно реперных точек выбирается N_0 дискретных точек с шагом дискретизации по истинной аномалии $\Delta\nu = 360^0 / N_0$, где значение N_0 определяется из следующих условий: 1) статистической значимости и репрезентативности выборки; 2) уровня надежности оценки фрактальной меры. В-третьих, на каждом i-ом шаге дискретизации по $\Delta\nu$ вычисляется расстояние r_i между соответствующими точками на орбитах. В-четвертых, после завершения предыдущего шага по всем дискретным точкам орбит определяются r_{min} и r_{max}, вычисляется $\Delta = |r_{max} - r_{min}|$, который разбивается на K подинтервалов. В-пятых, для заданного уровня геометрической близости двух орбит $r_{дов}$ (радиус сферы доверительности) вычисляется оценка уровня доверительности $p_{дов}$, который отражает долю точек $r_i \in \Delta$, для которых справедливо условие $r_i \le r_{дов}$. Оценка $p_{дов}$ является критерием и количественной мерой фрактальной природы близости двух орбит, т. е. для заданного $r_{дов}$ и с каким значением $p_{дов}$ можно считать связанными две орбиты. Для получения количественной оценки фрактальной меры на K подинтервалах множества точек $r_i \in \Delta$, учитывающей одновременно емкостные и информационные размерности фрактала, используется формула для оценки универсальной фрактальной размерности d_b, которая является синергией и обобщением обозначенных выше фрактальных размерностей. Она была получена одним из авторов, а содержательно-смысловое значение величины d_b более полно и конструктивно описано в [5]. Оценка значения величины d_b определяется следующим выражением

$$d_b = \lim_{\varepsilon \to 0} \frac{B(\varepsilon)}{\log(1/\varepsilon)} = \lim_{\varepsilon \to 0} \frac{\sum_{i=1}^{K} p_i \log \sum_{j=1}^{K} (1 - \rho_{ij}) p_j}{\log \varepsilon}, \qquad (1)$$

где p_i – вероятность попадания значения r_i в i-й подинтервал $\Delta = |r_{min} - r_{max}|$; ε – длина подинтервала для заданного разбиения интервала Δ; ρ_{ij} – рандомизированная метрика между центрами j-го и i-го подинтервалов; $B(\varepsilon)$ – B-энтропия.

Рандомизированная метрика ρ_{ij} определяется по следующей формуле

$$\rho_{ij} = |r_i - r_j| \ / \ |r|, \qquad (2)$$

где $|r_i - r_j|$ – это расстояние (геометрическое или информационное) между i-м и j-м подинтервалами; $|r|$ – длина интервала Δ. Для вычисления оценки емкостной фрактальной размерности множества $\{r_i\}$, которая отражает свойства фрактальной геометрии «дырявого» множества точек $r_i \in \Delta$, использовалась известная формула [6]

$$d_f = -\lim_{\varepsilon \to 0} \frac{\log N(\varepsilon)}{\log \varepsilon}, \qquad (3)$$

где $N(\varepsilon)$ – число покрытий множества точек $r_i \in \Delta$; ε – радиус сферы покрытия. Емкостная фрактальная

размерность d_f вводится и определяется как оценка меры нерегулярности топологической и геометрической структуры множества точек r_i.

Формулы (1) и (3) использовались для получения оценок критериев связанности двух орбит, которые описывают и отражают степень пространственной близости двух орбит в определенной пространственной окрестности $r_{дов}$ с заданным уровнем доверительности $p_{дов}$. Логическая схема получения оценок фрактальной меры связанности потока орбит относительно оптимальной опорной орбиты и нахождение такой орбиты в потоке выражается в виде следующего алгоритма.

Первый шаг – произвольно выбирается первая фиксированная орбита потока O_j ($j=1\div L$, где L – количество орбит потока) и вычисляются по-парные индексы связанностей I_k между ней и всеми другими орбитами O_k ($k=1\div L$, $j\neq k$, k – переменный индекс) потока по формулам (1), (2). Результат этого шага – множество индексов связанностей $K_j=\{I_k\}$ для всех орбит потока относительно выбранной опорной орбиты O_j. На этом шаге также определяются индексы связанностей I_m, которые вычисляются по формуле (3) и отражают свойства фрактальной геометрии близости двух орбит, образуя множество $G_j=\{I_m\}$.

На втором шаге вычисляется коэффициент связанности потока орбит $\{O_k\}$ относительно опорной орбиты O_j по формуле (1), т. е. вычисляется фрактальная размерность R_j множества K_j. Процедуры первого и второго шагов выполняются для всего множества орбит $\{O_j\}$, где j – индекс опорной орбиты. На последнем шаге алгоритма получаем множество $S_{орб}=\{R_j\}$, элементы которого показывают, насколько та или иная орбита может быть «центром» потока орбит, относительно которого наиболее плотно и компактно группируются орбиты. Условие выбора такого «центра» потока орбит определяется выражением

$$R_j \to \min_j . \qquad (4)$$

Основная посылка и смысл обозначенной логической схемы в информационных технологиях обработки и анализа состоят в том, чтобы дать математическое и логическое описание пространства возможных состояний наблюдаемых малых тел с учетом геометрических, динамических и информационных аспектов их эволюции. Для представления расположения тел в моменты их наблюдений и дальнейшей эволюции разработана математико-логическая схема перехода от орбитального описания положений тел к их геометрическому расположению в реальном пространстве, которая реализована в виде следующей процедуры.

Во-первых, локализуется область пространства (куб или прямоугольный параллелепипед), в котором находятся наблюдаемые объекты-тела. Во-вторых, выделяется элементарный объём разбиения этой области, т. е. область разбиения представляется в виде трёхмерной решётки, узел которой является идентификатором элементарного объёма. В-третьих, решается задача таксономии и классификации пространственного распределения малых тел на узлах трёхмерной решётки. Задача заключается в следующем: образует ли распределение тел в этом объёме регулярную либо нерегулярную пространственную структуру? Для этого использовался аппарат теории фрактальных размерностей и фрактальной геометрии [5–7]. Решением задачи является выделение объема, определение пространственной геометрии и распределения тел на узлах решётки, т. е. определяется, являются ли соответствующие подмножества узлов решётки фрактальными объектами или регулярными. В-четвертых, решалась задача пространственной кластеризации наблюдаемых объектов на узлах решетки: выделения фрактальных кластеров – перколяционный фрактал или фрактальный агрегат.

2.2 Фрагменты результатов обработки и анализа данных наблюдений

Для обработки и анализа данные астрономических наблюдений для малых тел солнечной системы были сгруппированы, исходя из опыта предыдущих исследований [8–11]. Классификация малых тел по группам, полученная, исходя из ограничений на значения элементов их кеплеровских орбит, показана в Таблице 1.

Иллюстрация и отражение результатов обработки и анализа потока астрономических данных на примере объектов-тел группы 7 в виде фрагмента приведены в Таблице 2. В столбцах 2, 3 и 4 приведены фрагменты (подмножества) для множеств G_j и K_j ($j=2$) и соответствующие их элементам значения B-энтропии. Значения элементов этих множеств вычислялись относительно второй орбиты. Как видно из Таблицы 2, значения всех элементов множества G_j больше единицы (условие геометрической регулярности – это равенство всех элементов единице), тем самым отражены фрактальные свойства пространственной структуры потока орбит группы 7. Логические аналогии для элементов множества K_j (условие однородности геометрии потока – это равенство всех элементов нулю) указывают на то, что «дырявая» пространственная геометрия потока траекторий обладает неоднородной плотностью. На основе элементов пятого столбца была определена орбита, которая является «центром притяжения» потока, т. е. относительно нее наиболее плотно сгруппированы орбиты тел группы. Элементы этой орбиты на момент наблюдений имеют следующие значения: a=2.211545 а.е., i=7.50626^0, e=0.5894668, ω=123.26392^0, Ω=313.24332^0. Этот этап обработки и анализа данных в логической цепочке информационных технологий позволяет получить оценки количественной меры фрактальной природы и фрактальной геометрии потока орбит рассматриваемой группы тел, а также определить центр потока. Вторым этапом в логической цепочке информационных технологий обработки и анализа являются пространственное представление и

описание распределения малых тел исследуемой группы в локальной области гелиоцентрической прямоугольной системы координат. Гелиоцентрическая система координат *XOYZ* определялась с началом ***O*** в барицентре Солнечной системы. Плоскость *XOY* – плоскость эклиптики. Ось *OX* направлена в точку ♈ (точка весеннего равноденствия). Ось *OY* перпендикулярна ей. Ось ***OZ*** выбрана так, чтобы система векторов *OX, OY, OZ* образовала правую тройку.

Проиллюстрируем результаты, полученные на этом этапе, на примере объектов-тел группы 1. Разбиение объема, в котором локализованы тела на момент наблюдения, определяется ограниченной трехмерной решеткой Z^3 следующего масштаба: по оси X – 15 узлов, по оси Y – 14 узлов, по оси Z – 12 узлов. Процедура пространственной кластеризации на решетке Z^3 позволила получить следующие результаты.

Во-первых, был выделен 31 кластер: один кластер типа перколяционного фрактала и тридцать типа фрактального агрегата. Размер перколяционного фрактала составлял 473 узла, а размеры фрактальных агрегатов варьировались от одного до нескольких узлов.

Во-вторых, перколяция обнаружена в плоскости *XOY*.

В-третьих, степень заполнения узлов решетки Z^3 составляла в пределах двадцати одного процента: около восемнадцати процентов занимает перколяционный фрактал, а остальное – фрактальные агрегаты.

Аналогичные расчеты для тел группы 7 с теми же размерами решетки Z^3 показали следующие результаты: не было обнаружено ни одного перколяционного фрактала, а только кластеры типа фрактального агрегата в количестве 31 (размеры от 1 до 9 узлов), т. е. объекты этой группы не образуют компактные пространственно-протяженные образования. Процессы перколяции малых тел на узлах решетки Z^3 достаточно полно отражают и описывают пространственную и эволюционную связанность этих объектов на их орбитах. Эти процессы наиболее характерны для объектов группы 7, указывая на то, что пространственно-временная геометрия этих объектов в пространстве возможных состояний имеет фрактальную природу, тем самым определяя тип их фрактальной динамики. Для этого типа динамики характерна наиболее нерегулярная пространственно-временная фрактальная геометрия. Такой геометрией с небольшими пространственно-времен-ными масштабами обладают малые тела типа метеорных тел и ряда других. Фрактальная агрегация в пространстве возможных состояний наиболее ярко проявляется в других группах малых тел. Такие фрактальные структуры отражают другой тип фрактальной динамики, проявляемой в различных пространственно-временных масштабах фрактальной геометрии по-разному: при больших – это более регулярный, а при небольших – менее регулярный тип динамики.

Таблица 1 Классификация малых тел по элементам кеплеровских орбит

Группа 1	*Главный пояс астероидов: e<1/3; i<20°; 2.1<a<3.5 а. е.*
Группа 2	*Короткопериодические кометы и метеорные тела (включая астероиды группы Аполлона – Амура):* 1/3<e<0.95*; i<30°; a<15 а. е.*
Группа 3	*Долгопериодические кометы и метеорные тела: e>0.95; i – случайное; a>15 а. е.*
Группа 4	*Троянцы (захваченные Юпитером и колеблющиеся относительно его передней и задней лагранжевых точек либрации): a≈5.2 а. е.*
Группа 5	*Астероиды группы Гильды: e ≈0.2; i≈10°; a ≈3.95 а. е.*
Группа 6	*Астероиды группы Венгрии: e≈0.1; i≈25°; a≈1.9 а. е.*
Группа 7	*Малые тела, орбиты которых пересекают орбиту Земли в окрестности радиуса сферы ее влияния.*

Таблица 2 Значения для элементов множеств G_j и K_j (j=2) и соответствующие этим элементам значения *B*-энтропии [5]

№ №	B-энтропия	Элементы множества G_j	Элементы множества K_j	Элементы множества $S_{орб}$
1.	0.4750731	1.2494696	0.20632162	0.026309
2.	0	1	0	0.036631
3.	0.4301281	1.2460074	0.18680227	0.033831
4.	0.4807579	1.2096293	0.20879053	0.025871
5.	0.5052235	1.1657263	0.21941579	0.028186
6.	0.4058040	1.2416164	0.17623847	0.036397
7.	0.4985102	1.1521853	0.21650024	0.026056
8.	0.4528685	1.2466675	0.19667829	0.022999
9.	0.4474228	1.1198819	0.19431325	0.030783
10	0.4823662	1.2204674	0.20948901	0.026293
11	0.4808495	1.2688664	0.20883028	0.022580
12	0.5008954	1.2177231	0.21753612	0.031327

3 Выводы и некоторые обобщения

Результаты обработки и анализа данных наблюдений для групп малых тел, обозначенных выше, позволяют сделать ряд выводов и обобщений следующего характера. Во-первых, фрактальные методы в информационных технологиях обработки и анализа больших потоков астрономических данных на основе логических схем когнитивной аналитики раскодирования сокрытой в них информации являются перспективной и уникальной парадигмой в области разработки информационных технологий нового поколения для широкого класса задач не только современной астрономии. Во-вторых, потоки данных астрономических наблюдений можно обрабатывать, используя различные процессы и методы теории фракталов и генетических данных как

для получения совокупностей и популяций выборочных данных, так и для их анализа. Эти методы и процессы отражают и определяют особенности получаемых оценок фрактальных мер и размерностей, а также область применения выводов, которые можно сделать на основе этих данных. В этом случае используются два типа выборочности – генетическая и статистическая. Статистическая выборочность связана с определением пространственных масштабов решетки Z^3, а генетическая – с распределением информации и объектов на узлах этой решетки.

В широком аспекте фундаментальных астрономических исследований проблемы образования и эволюции планетных систем результаты данной работы впервые позволили показать, как и в чем проявляется синергия геометрии пространственной структуры и динамической эволюции объектов обозначенных систем и как это можно описать и объяснить в рамках фрактальной парадигмы. Можно ли провести такие аналогии в рамках традиционных моделей, алгоритмов, схем и др.? Если да, то необходимо показать результаты обозначенных аналогий и сформулировать тренды их теоретического развития и практического продолжения.

Прикладные аспекты результатов работы тесно связаны с решением задач астероидно-кометно-метеорной безопасности и проблемой космического мусора. С одной стороны, предложены методы фрактальной теории решения сложных нелинейных задач обработки, анализа и интерпретации результатов динамической эволюции объектов космического пространства с нерегулярной пространственно-вре-менной фрактальной геометрией. С другой стороны, разработана и реализована новая IT-технология в тренде DAMDID обработки, анализа и классификации орбитальных данных малых тел Солнечной системы (для программной реализации IT-технологий использованы данные с порталов MPC (www.cfa.harvard.edu) и NASA (www.nasa.gov)).

Литература

[1] Гусева, И.С., Лих, Ю.С.: Статистический анализ орбит комет. Известия ГАО РАН, 220, сс. 219-224 (2012)

[2] Кочетова, О.М., Кузнецов, В.Б., Медведев, Ю.Д., Шор, В.А.: Каталог элементов орбит нумерованных астероидов ИПА РАН. Известия ГАО РАН, 220, сс. 255-258 (2012)

[3] Малкин, З.М.: Некоторые результаты статистического анализа определений галактического расстояния Солнца. Известия ГАО РАН, 220, сс. 401-406 (2012)

[4] Брюно, А.Д., Варин, В.П.: О распределении астероидов по средним движениям. Астрономический вестник, 45 (1), сс. 334-340 (2011)

[5] Мышев, А.В.: Метрологическая теория динамики взаимодействующих объектов в информационном поле нейросети и нейрона. Информационные технологии, 4, сс. 52-63 (2012)

[6] Павлов, А.Н., Онищенко, В.С.: Мультифрактальный анализ сложных сигналов. УФН, 7 (8), сс. 859-876 (2007)

[7] Федер, Е.: Фракталы. М.: Мир (1991)

[8] Емельяненко, В.В., Нароенков, С.А., Шустов, Б.М.: Распределение околоземных объектов. Астрономический вестник, 45 (6), сс. 512-517 (2011)

[9] Гафтонюк, Н.М., Горькавый, Н.Н.: Астероиды со спутниками: анализ наблюдательных данных. Астрономический вестник, 47 (3), сс. 213-220 (2013)

[10] Нароенков, С.А.: Хранение и обработка астрометрических и фотометрических данных об АЗС: настоящее и будущее в России. Космические исследования, 48 (5), сс. 467-470 (2010)

[11] Альвен, Х., Аррениус, Г.: Эволюция Солнечной системы. М.: Мир (1979)

On the Problem of Multi-word Term Extraction from a Domain-specific Document Collection

© M.S. Karyaeva © V.A. Sokolov

Yaroslavl State University,
Yaroslavl, Russia

mari.karyaeva@gmail.com valery-sokolov@yandex.ru

Abstract. This paper presents some methods for multi-word term extraction from a domain-specific collection of documents. We develop the approach, based on topic modeling and word2vec algorithms. The domain-specific area is poetics.

Keywords: TermExtraction, TopicModel, Word2Vec, Thesaurus.

1 Introduction

The development of a high quality terminological database is the first step to create a comprehensive domain-specific thesaurus. The quality of semantic relation between terms in the thesaurus depends to a great extent upon the high coverage of the terminological database. The most common way to solve this problem is to enrich the terminological database with time consuming manual terms extraction by domain assessors. However, this approach may cause a coverage loss. Automatic terms extraction can be exploited for developing domain-specific thesauri and ontology, entity and fact extraction, information retrieval. The observation shows, that single-word terms appear frequently, but automatic single-word term extraction is not enough for developing high quality thesauri.

This paper focuses on multi-word terms. The method of multi-word term extraction includes two steps:

1) To extract a candidate and detect the candidate as a term;
2) To define the candidate belongs to a domain-specific area.

The paper is organized as follows: Section 2 reviews the related work. In Section 3 we give an overview of existing methods for terms extraction. Section 4 presents experimental results. Finally, we conclude with a summary in Section 5.

2 Related work

The automatic terms extraction task has been applied to different domains, such as medicine [1, 2], banking [3], mathematics [4], etc. In this article we provide an overview of basic methods for automatic extraction of terms and indicate papers using them. In a large list of papers the process of automation does not appear as a final method in the construction of the terminology database. Researchers are interested in getting a universal method with the highest percentage of probability of extracting terms.

Proceedings of the XIX International Conference "Data Analytics and Management in Data Intensive Domains" (DAMDID/RCDL'2017), Moscow, Russia, October 10–13, 2017

Kiseljov et al. [5] present the method of semantic relations based on a structure of dictionaries, that the definition of a concept contains a candidate, denoting the main concept in relation to the word being determined. We decide to apply the structure to extract terms increased the length of terms from 1-word to multi-word terms.

In addition, the vector representation [13] for words, which is the state-of-the-art word embeddings for data experiments. The word embeddings is applied in a wide range of natural language processing tasks. The text corpora is fed to the word2vec algorithm, after learning the model with certain parameters, vector representations are formed as the output with reflection of words semantics. We applyword2vec algorithm as an indicator of reliability in the automatic terms extraction.

This overview research would not be complete if we did not consider the topic modeling, so we carry out the experiment, relying on papers [10, 11].

3 Overviewofcurrentmethods

*Linguistic methods.*The accurate development of linguistic patterns leads to high accuracy term extraction. Templates can be designed with a number of features regarding morphology, punctuation, and the rules of sentence-construction, for instance, the development of templates for extracting terms from explanatory dictionaries [5]. Terms can be extracted with high accuracy by using this method, but the coverage will remain low due to the lack of template flexibility.

Statistical methods. These methods rely on frequency characteristics, which denote the relevance of the term in a document by counting its occurrence. TF-IDF, which determines weights for each word in document collections, is a main measure implementing in statistical methods. To increase the quality of the results, stop words (prepositions, conjunctions, etc.), widely-used words, words with a low frequency can be removed from the potential list of candidates.

In 1996, the paper [6] introduced the concept of termhood, indicating the connection degree of the extracted candidate with the selected domain area. C-Value measurement has been introduced for the extraction of multi-word terms [7] and it is designed for getting the termhood among potential candidates. The

application of C-Value measurement (in a direct and modified form) demonstrated positive results in a biomedical field [8].

Methods based on machine learning algorithms. In this case Machine learning algorithms for term extraction intersect the classification task, in particular, the separation of words and phrases into two groups: term-candidates and other words or phrases.

Paper [9] describes a successful example of using machine learning algorithms for term extraction task. The authors conducted an experiment with different models, though Begging classification (Bootstrap aggregating) with decision trees had been chosen as the crucial algorithm.

Methods based on topic modeling. Topic models are designed to identify the topic groups presented in the document collection, as well as the extraction of a list of terms belonging to each topic group.

Meanwhile, the systematic regularity is observed: synonyms are distributed in the same group, and homonyms (words different in meaning but similar in spelling) are distributed in different topic groups with a high probability because of the difference in the contextual environment. Some topic models such as LDA, PLSA represent a document like a set of words without semantic relations and without word order that is "BOW" (bag of words). However, paper [10] presents N-grams topic model that can discover phrases.

4 Experimental results

First step: manual labeling

In this paper we deal with a specific domain – poetics. There are several reasons, why the area has been chosen. Firstly, this research is the subtask of total project of development Russian poetics system, which includes a thesaurus and an intelligence block for poems analysis. Secondly, domain experts work on this task, so they can evaluate with a high precision the automatic results. Finally, we are pioneers in that area and we cannot compare our results with other because the area is non-mining already.

At the first stage of work a list of poetics terms was compiled manually by domain assessors. This list contains 1 544 unique domain-specific terms. Part-of-speech tagging was performed to identify phrases patterns. We have collected data in statistics, using the manually created list of terms and have defined the most frequently occurring multi-word terms (A = Adjective, S = Substantive, ADV = Adverb, PR = Pretext), see Table 1:

Table 1 Statistics of some phrases patterns

Scheme	Number of occurrences	Example
A_S	323	Французская баллада
S_A	70	Рифмовка кольцевая
S_S	64	Растяжение слогов
S_A_S	14	Дериват стихотворного размера
A_A_S	13	Национальная стиховая культура
A_S_S	13	Эпистолярная форма языка
ADV_A_S	11	Произвольно ударный слог
S_PR_S	9	Роман в стихах
S_ADV_A	7	Слог обязательно ударный
S_S_A	4	Модель стиха языковая

Second step: linguistic method

The next step included extracting terms from domain-specific dictionaries using linguistic methods. A list of domain-specific sources for applying linguistic method:

- Brief Literary Encyclopedia: In 11 volumes (BLE)
- Literary Encyclopedia: In 11 volumes (LE)
- Dictionary of Literary Terms: In 2 volumes (DLT)
- Kwiatkowski. Glossary of Poetic Terms. (GPT by Kwiatkowski)

Table 2 The comparison of manual labeling and linguistic method

Source title	The intersection of the list of terms and a source	Total terms in a source	Correlation
BLE	571	15 228	3,7%
LE	401	4 782	8,3%
DLT	286	739	38%
GPT	601	673	89%
Do not occur in any source	587	--	--
Occur in all sources	128	--	--

We iteratively developed series of patterns for a single word and multi-word terms extraction by using the pattern "term - definition". Extracted terms do not always belong to the selected topic, because a term can be assigned to famous figures (poets, playwrights, critics, etc.) and other words closely related to the topic. Table 2 presents the comparison of manual labeling and linguistic method: it is the correlation of terms extracted from the documents with terms that had been extracted manually by domain experts. Thus, Kwiatkowski glossary contains the highest number of terminology occurrences. Furthermore, 587 terms do not appear in any source presented and 128 terms appear in each presented source.

Third step: automatic extraction (in progress)

The current stage is divided into two parts:

1. Exploiting the topic model for terms extraction;
2. Checking the relevance of candidates applying for associative words getting from the Word2Vec.

Exploiting the topic model to extract single-word

terms shows proper results [11]. Three algorithms were chosen as topic models: TNG [10], PDLDA [12] и PLSA-SIM [11].

To research the poetics area, the experts selected 31 domain-specific sources, converting them into electronic form. The sources were applied to test models (the list of sources is given in Appendix). Morphological analysis of text collections and words lemmatization were performed at the stage of pre-processing. The templates obtained in the first step were used as extraction patterns. To improve the quality of term extraction, a list of stop-words was developed. Word2Vec models were trained to confirm the candidates [13].

As a result, we obtained a vector representation of the main word and associative words with a measure of similarity, which is presented in cosine similarity between the vectors of a main word and associate words.

Associative words can be presented as context words describing the main word, as well as words that have semantic relationships with the main word.

Training of Word2Vec model was conducted on the Russian Wikipedia data because it includes a connected terminology graph of domain-specific area. The dump of the Russian Wikipedia articles is 16 GB.

Further, associate words were built for each candidate, obtained in step 1. The associative words are indicators of an adoption of the candidate in the list of domain terms. An indicative measure is positive when the terms from list of terms or new added terms have an exact match with a candidate.

4 Conclusion

This article is devoted to the automatic terms extraction from text corpora. We experimentally show that it is possible to combine linguistic, statistical and methods based on machine learning to enhance automatic terms extraction task. Exploiting a combination of linguistic and statistical methods provides the basis to obtain a better quality of results in machine learning algorithms. We have developed a hypothesis of evaluation extracted candidates using associative words obtained from word2vec.

The candidates were obtained by applying the topic modeling algorithms. In the future we plan to conduct expert estimation and compare the results with other algorithms to produce meaningful findings.

Acknowledgment

This work was supported by the Russian Foundation for Basic Research (RFBR), project № 16-07-01180.

References

[1] Abacha A. B., Zweigenbaum P. Automatic extraction of semantic relations between medical entities: a rule based approach //Journal of biomedical semantics. – 2011. – Т. 2. – №5. – С.1

[2] Lossio-Ventura J. A. et al. BioTex: A system for biomedical terminology extraction, ranking, and validation //Proceedings of the 2014 International Conference on Posters & Demonstrations Track-Volume 1272. – CEUR-WS. org, 2014. – С. 157-160

[3] Dobrov B. V., Loukachevitch N. V. Multiple Evidence for Term Extraction in Broad Domains //RANLP. – 2011. – С. 710-715

[4] Stoykova V., Petkova E. Automatic extraction of mathematical terms for precalculus //Procedia Technology. – 2012. – Т. 1. – С. 464-468

[5] Kiseljov Ju. A., Porshnev S. V., Muhin M. Ju. Metod izvlechenija rodovidovyh otnoshenij mezhdu sushhestvitel'nymi iz opredelenij tolkovyh slovarej // Programmnaja inzhenerija. Vyp 10, 2015 – S. 38—48

[6] Kageura K., Umino B. Methods of automatic term recognition: A review //Terminology. – 1996. – Т. 3. – №. 2. – С. 259-289

[7] Frantzi K., Ananiadou S., Mima H. Automatic recognition of multi-word terms:. the c-value/nc-value method //International Journal on Digital Libraries. – 2000. – Т. 3. – №. 2. – С. 115-130.

[8] Lossio-Ventura J. A. et al. Combining c-value and keyword extraction methods for biomedical terms extraction //LBM: Languages in Biology and Medicine. – 2013

[9] Lopez P., Romary L. HUMB: Automatic key term extraction from scientific articles in GROBID //Proceedings of the 5th international workshop on semantic evaluation. – Association for Computational Linguistics, 2010. – С. 248-251

[10] Wang X., McCallum A., Wei X. Topical n-grams: Phrase and topic discovery, with an application to information retrieval // Proceedings of the 7th IEEE International Conference on Data Mining. – 2007. – P. 697-702

[11] Nokel M., Loukachevitch N. A Method of Accounting Bigrams in Topic Models //Proceedings of NAACL-HLT. – 2015. – С. 1-9.

[12] Lindsey R. V., Headden III W. P., Stipicevic M. J. A phrase-discovering topic model using hierarchical pitman-yor processes //Proceedings of the 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning. – Association for Computational Linguistics, 2012. – С. 214-222

[13] Mikolov T., Yih W., Zweig G. Linguistic Regularities in Continuous Space Word Representations //HLT-NAACL. – 2013. – С. 746-751

APPENDIX

The list of the main scientific poetics sources

[1] Гаспаров М. Л. Очерк истории русского стиха: Метрика; Ритмика; Рифма; Строфика. 2-е изд., доп. М.: Фортуна Лимитед, 2000

[2] Гаспаров М. Л. Русский стих начала XX века в комментариях. 2-е изд., доп. М.: Фортуна Лимитед, 2001

[3] Гаспаров М. Л. Очерк истории европейского стиха. 2-е изд., доп. М.: Фортуна Лимитед, 2003

[4] Гаспаров М. Л. Метр и смысл: Об одном из механизмов культурной памяти. М.: РГГУ, 1999. Переизд., М.: Фортуна Эл, 2012
[5] Гаспаров М. Л. Избранные труды. М.: Языки рус. культуры, 1997. Т. I—III. М.: Языки слав. культуры, 2014. Т. IV
[6] Жирмунский В. М. Теория стиха. Л.: Сов. писатель, 1975
[7] Жирмунский В. М. Теория литературы. Поэтика. Стилистика. Л.: Наука, 1977
[8] Илюшин А. А. Русское стихосложение. [2-е изд., доп.] М.: Высш. шк., 2004
[9] Лотман Ю. М. Анализ поэтического текста: Структура стиха. Л.: Просвещение, 1972
[10] Тарановский К. Русские двусложные размеры. Статьи о стихе / Пер. с серб. В. В. Сонькина. М.: Языки слав. культуры, 2010
[11] Тарановский К. О поэзии и поэтике. М.: Языки рус. культуры, 2000
[12] Тимофеев Л. И. Очерки теории и истории русского стиха. М.: Гос. изд-во художественной литературы, 1958
[13] Томашевский Б. Русское стихосложение. Метрика. Пг.: Academia, 1923. (Вопросы поэтики; Вып. II)
[14] Томашевский Б.Стих и ритм: Методологические заметки // Поэтика: Временник Отдела словесных искусств Гос. ин-та истории искусств. Л.: Academia, 1928. Вып. IV
[15] Томашевский Б. В. Строфика Пушкина // Пушкин: Исследования и материалы. М.; Л.: Изд-во АН СССР, 1958. Т. II
[16] Томашевский Б. В. Стих и язык. М.: Гос. изд-во художественной литературы, 1959
[17] Тынянов Ю. Н. Проблема стихотворного языка [1924] // Тынянов Ю. Н. Литературная эволюция. Избранные труды. М.: Аграф, 2002
[18] Холшевников В. Е. Основы стиховедения. Русское стихосложение: Учебное пособие: Для студентов филол. фак. 4-е изд., испр. и доп. М.: Academia; СПб.: Филол. фак. СПбГУ, 2002
[19] Шапир М. И.Universumversus: Язык — стих — смысл в русской поэзии XVIII—XX веков. М.: Языки рус. культуры, 2000. Кн. 1
[20] Шапир М. И.Universumversus: Язык — стих — смысл в русской поэзии XVIII—XX веков. М.: Языки рус. культуры, 2015. Кн. 2
[21] Шенгели Г. Трактат о русском стихе. Ч. I: Органическая метрика. Изд. 2-е, перераб. М.; Пг.: Гос. изд-во, 1923
[22] Шенгели Г. Техника стиха. [Изд. 4-е]. М.: ГИХЛ, 1960
[23] Эйхенбаум Б. Мелодика русского лирического стиха. Пб.: ОПОЯЗ, 1922
[24] Эйхенбаум Б. М. Лермонтов: Опыт историко-литературной оценки. Л.: Гос. изд-во, 1924
[25] Эйхенбаум Б. О поэзии. Л.: Сов. писатель, 1969
[26] Якобсон Р. Работы по поэтике. М.: Прогресс, 1987
[27] Arspoetica. М.: ГАХН, 1928. Вып. II: Стих и проза. (Труды ГАХН. Лит. секция; Вып. 2)
[28] Теория стиха. Л.: Наука, 1968
[29] Исследования по теории стиха. Л.: Наука, 1978
[30] Проблемы теории стиха. Л.: Наука, 1984
[31] Русское стихосложение: Традиции и проблемы развития. М.: Наука,1985

Анализ и визуализация международного научного сотрудничества на основе научных публикаций

© А.И. Майсурадзе © Е.Ю. Ечкина

Московский государственный университет имени М. В. Ломоносова,
Москва, Россия

maysuradze@cs.msu.ru ejane@cs.msu.ru

Аннотация. В настоящее время активно развивается бизнес-аналитика науки. В работе рассмотрена задача анализа и визуализации множественного взаимодействия разных единиц анализа. Исходная информация взята из индекса научных публикаций, что создает интенсивный поток данных. Соответственно, единицами анализа являются акторы, характеризующие публикацию: авторы, страны, научные центры, ключевые слова. Каждая публикация рассматривается как событие, характеризуемое множеством акторов. Представлены методы и результаты иерархической кластеризации акторов, приведены примеры визуализации.

Ключевые слова: наукометрия, индекс научных публикаций, неатомарные данные, кластеризация, визуализация.

Analysis and Visualization of International Scientific Cooperation Based on the Scientific Publication Index

© A. Maysuradze © E. Echkina

Lomonosov Moscow State University,
Moscow, Russia

maysuradze@cs.msu.ru ejane@cs.msu.ru

Abstract. Currently, the business analytics of science is actively developing. We consider the problem of analysis and visualization of multiple interactions of different units of analysis. The initial information is taken from a scientific publication index, which is data intensive. Accordingly, units of analysis are the actors characterizing a publication: authors, countries, research centers, keywords. Each publication is regarded as an event characterized by a set of actors. We present some methods and results of hierarchical clustering of actors, and provide some examples of visualization.

Keywords: scientometrics, scientific publication index, non-atomic values, multi-valued attributes, cluster analysis, visualization.

1 Введение

Предметная область данного исследования относится к наукометрии и бизнес-аналитике науки. Содержательная задача состоит в анализе взаимодействия различных акторов публикационной активности. Математическая проблема состоит в кластеризации категорий по выборке объектов, имеющих мультикатегориальное описание. Подчеркнем, что кластеризуются не исходные объекты, а описывающие их категории. Результаты указанной кластеризации нужно визуализировать.

Библиографическая информация о научных публикациях довольно широко используется при определении и расчете показателей результативности различных акторов научной деятельности (ученые, научные центры и т. д.). Можно сказать, что наука стала предметной областью с интенсивным использованием бизнес-интеллекта (business intelligence) – всевозможных инструментов анализа данных, повышающих эффективность бизнес-аналитики (business analytics) и скорость принятия решений. Более того, такое использование научных публикаций повлекло изменение традиций оформления их выходной информации: сегодня публикация не только содержит научную составляющую, но и является богатым источником сведений об индивидуальной работе и взаимодействии акторов. Сложность такого взаимодействия приводит к тому, что с точки зрения

Труды XIX Международной конференции «Аналитика и управление данными в областях с интенсивным использованием данных» (DAMDID/ RCDL'2017), Москва, Россия, 10–13 октября 2017 года

традиционных моделей данных описание каждой публикации имеет довольно неудобную структуру. Если пытаться описать публикацию набором атрибутов, то целому ряду важнейших из них придется позволить принимать множественные (неатомарные) значения.

Основным источником библиографической информации в наши дни являются реферативные базы публикаций, одновременно служащие индексами научного цитирования. В данной работе использованы данные из базы Web of Science [5]. В последние годы эта база ежеквартально растет примерно на 300 тысяч публикаций, что создает достаточно интенсивный источник данных.

В настоящее время наукометрические инструменты, встроенные в платформы индексов цитирования, в первую очередь ориентированы на оценку результативности отдельных акторов научной деятельности. Современный социально-экономический подход к исследованию науки порождает более сложные аналитические запросы: требуется анализировать попарное и групповое взаимодействие акторов.

В данной работе содержательная задача анализа взаимодействия групп акторов конкретизирована как задача выделения сообществ акторов, схожим образом участвующих в публикационной деятельности. Научная публикация рассматривается как событие, с которым связано множество акторов. Сходство будет определяться по распределению вероятности оказаться в одной публикации с другим актором.

Вообще говоря, рассматриваемый ниже подход применим к любым системам, в которых объекты характеризуются множествами категорий и требуется сгруппировать последние. Для публикаций такими атрибутами с множественными значениями являются авторы, научные центры, страны, ключевые слова, т. е. термин «актор» можно понимать в широком смысле категории измерения.

Идея подхода состоит в следующем. От множества публикаций мы переходим к мультиграфу попарных взаимодействий акторов. На мультиграфе решается задача выявления плоских сообществ акторов. На основе модели выявления плоских сообществ решается задача выявления иерархии сообществ однотипных акторов. Далее используются методы визуализации плоских сообществ и иерархии сообществ на мультиграфе. Использование информации о сообществах позволяет существенно повысить интерпретируемость раскладки и раскраски мультиграфа.

Дальнейшая структура статьи соответствует перечисленным выше этапам предлагаемого подхода.

2 Мультиграф акторов

В исходных данных объектами являются события (публикации), каждое из которых связано со множеством категорий (акторов). Мы хотим, чтобы объектами стали категории.

Известной общей идеей такого преобразования является построение сети соавторства. Мы предлагаем строить мультиграф попарного взаимодействия акторов. Термин мультиграф означает, что между парой акторов может быть много ребер.

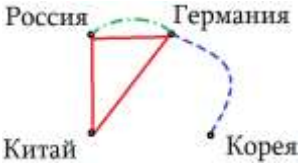


Рисунок 1 Мультиграф взаимодействия стран по трем публикациям

Вершинами мультиграфа являются категории (акторы). Каждая публикация для каждой пары связанных с ней акторов добавляет в мультиграф соответствующее ребро. Если публикация связана с n акторами, то в мультиграф добавляется $n(n–1)/2$ ребер. На рис. 1 приведен пример такого мультиграфа взаимодействия стран для трех публикаций: {Россия, Германия, Китай}, {Россия, Германия}, {Германия, Корея}.

Разумеется, в программной реализации просто рассчитывается количество совместных публикаций для каждой пары стран. Здесь мы говорим о мультиграфе, а не о взвешенном графе, поскольку ниже собираемся использовать генеративные модели, которые будут генерировать отдельные ребра, а не их веса.

Существуют традиционные методы раскладки и непосредственной визуализации графов и мультиграфов. К сожалению, результаты такой визуализации редко удовлетворяют экспертов в предметной области. Типичный пример визуализации графа сотрудничества стран можно видеть на рис. 2: никаких явных групп стран на первый взгляд не заметно.

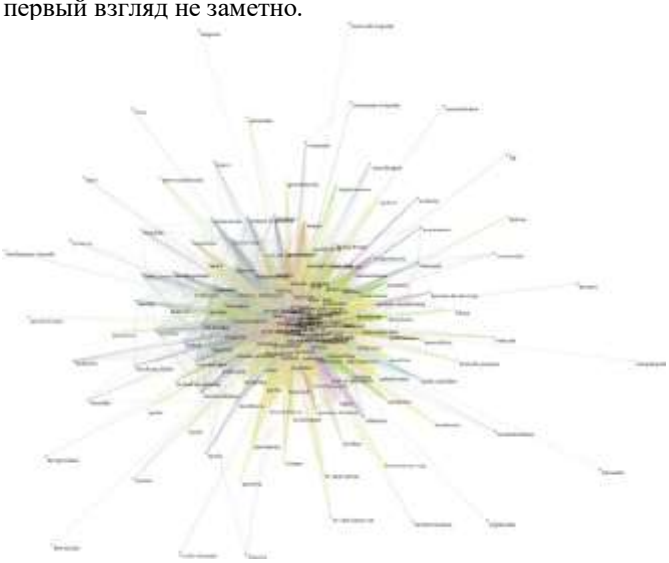


Рисунок 2 Пример распространенного, хотя и плохо интерпретируемого способа визуализации графа соавторства

Соответственно, требуется явно выделить структуру сообществ акторов, прежде чем визуализировать граф. Кроме того, такая структура сообществ является самоценным результатом даже без какой-либо визуализации графа соавторства. Если в качестве акторов выступают объекты, уже имеющие традиционную визуализацию, например, страны на карте, то структура сообществ может использоваться при визуализации для цветового кодирования.

Отметим, что в данной работе описание событий множествами событий не рассматривается. В частности, для публикаций не рассматриваются библиографические ссылки.

3 Плоская кластеризация

Метод плоской кластеризации вершин мультиграфа предлагается построить на базе генеративной модели, которая сразу будет включать в себя разбиение вершин на блоки как параметр. Тогда кластеризация произойдет при настройке параметров такой модели по исходным данным: нужно так выбрать блоки вершин, чтобы исходный граф с наибольшим правдоподобием генерировался моделью. В математической литературе это называется байесовским выводом параметров модели по данным: единственный наблюдаемый нами мультиграф воспринимается как свидетельство, подтверждающее гипотезу.

Для реализации данной идеи возьмем один из простых генеративных процессов, использующий блоки вершин, а именно, стохастическую блоковую модель [3]. Модель использует матрицу смежности разбиения, которую строит по матрице смежности вершин и разбиению. Ребра генерируются по матрице смежности разбиения. В такой модели вершины графа попадают в один блок, если схожи вероятности их соединения с другими вершинами.

Традиционной проблемой байесовского вывода является необходимость внешнего задания некоторых распределений. В классической стохастической блоковой модели [3] предполагается, что ребра распределены равномерно в каждой группе, а степени вершин в каждой группе практически совпадают. Такие предположения нельзя считать приемлемыми в нашем случае: трудно ожидать, что все страны выпустят примерно одинаковое число публикаций.

Улучшенная стохастическая блоковая модель вводит степени вершин как скрытые переменные. Процедура настройки такой модели имеет вычислительную сложность $O(V \ln 2V)$, где V – число категорий. От количества публикаций зависит сложность построения мультиграфа, но не сложность кластеризации категорий. Вывод осуществляется посредством семплирования из марковских цепей (Markov chain Monte Carlo [1]).

4 Иерархическая кластеризация

Подход, использованный для плоской кластеризации, позволяет пойти дальше и построить вложенную стохастическую блоковую модель. Это «стопка» из улучшенных стохастических блоковых моделей, причем каждый следующий слой задает свою априорную вероятность как апостериорную вероятность с предыдущего слоя [4]. В чем-то это напоминает агломеративную кластеризацию. Блоки следующего слоя как бы состоят из блоков предыдущего слоя. Но во всех слоях число ребер одно и тоже, определяемое исходным мультиграфом.

В итоге получается иерархическая структура сообществ однотипных акторов, например, группы групп стран. Количество слоев и количество блоков в каждом слое определяются автоматически.

5 Результаты экспериментов

Предложенные процедуры были опробованы на выборке публикаций из индекса научного цитирования Web of Science [5]. Были применены библиотеки Boost Graph и graph-tool. В экспериментах была использована одна из ежеквартальных рассылок данных подписчикам этой системы со следующими характеристиками: 364221 публикация из 4386 источников («журналов»), 398972 автора, 100671 ключевое слово (после нормализации), 69093 научных центров из 195 стран.

В качестве метода раскладки вершин и раскраски ребер по уже построенной иерархии сообществ использовался подход из [2], который показывает и сообщества, и ребра между акторами.

Для широкого круга читателей легче всего воспринимать результаты по группированию стран и ключевых слов, которые и приведены ниже.

На рис. 3 изображена политическая карта мира, на которой страны раскрашены в соответствии с кластеризацией на нижнем уровне. Более крупные блоки не обозначены. Хотя этот способ наиболее удобен для неподготовленной широкой аудитории, он передает лишь часть полученных результатов.

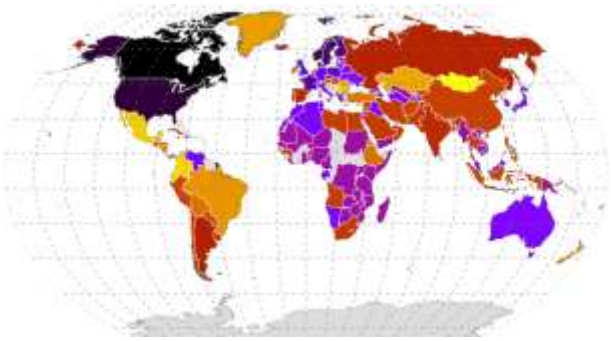


Рисунок 3 Группы стран на нижнем уровне кластеризации

На рис. 4 можно видеть всю иерархическую структуру сообществ стран. Кластеры разных уровней показаны и расположением вершин, и раскраской ребер, и траекториями ребер. На высоких

слоях структура сообществ отражает деление стран на научных лидеров в мире и регионах.

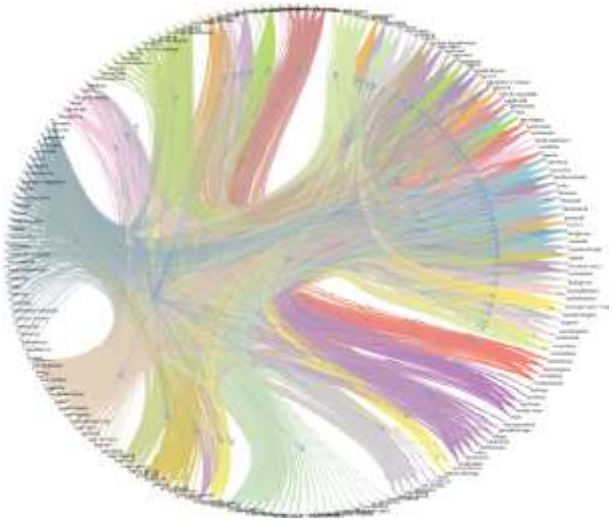


Рисунок 4 Полная иерархическая структура сообществ стран

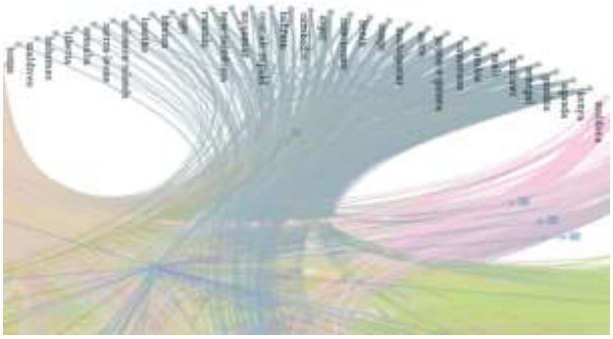


Рисунок 5 Фрагмент иерархической структура сообществ стран

На рис. 5 представлен фрагмент рис. 4 в надежде, что читатели смогут увидеть элементы визуализации дерева кластеров. Разумеется, в программной системе это векторная графика, которая может быть произвольно увеличена.

Тем же методом были кластеризованы и визуализированы ключевые слова. Представляется, что в тексте данного формата читателям будет удобнее увидеть перечень слов, чем картинку. Вот пример одного из кластеров: disease, infection, syndrome, resistance, serum, insulin, weight, glucose …

Время работы метода в основном определяется количеством акторов. Например, представленные выше иллюстрации на обычном ноутбуке были построены примерно за 30 секунд (кластеризация и визуализация). Отметим, что есть ощутимые затраты времени на отрисовку уже созданных векторных изображений стандартными программами, например, после сохранения векторных изображений в файл формата pdf.

6 Заключение

Предложен способ анализа и визуализации множества событий, характеризуемых множествами категорий. В результате выявляется и визуализируется иерархическая структура категорий. Невысокая вычислительная сложность метода позволяет применять его к достаточно обширным коллекциям событий.

Метод опробован на библиографической информации из реферативной базы публикаций Web of Science. Эксперты в предметной области соглашаются с результатами данного метода группирования стран, довольно легко пользуются полученными диаграммами, однако продолжается разработка новых специализированных методов визуализации такой информации.

В реальности взаимодействие акторов имеет более сложную структуру, чем строгая иерархия. Кроме того, публикация связывает акторов разных типов – например, научные центры и области знаний, – соответственно, модель должна быть гетерогенной. Продолжается работа по созданию таких моделей и методов их настройки по данным.

Благодарности

Работа выполнена при частичной поддержке РФФИ, проекты 15-07-09214, 16-57-45054, 16-01-00196. Авторы выражают благодарность бакалавру факультета ВМК МГУ имени М.В. Ломоносова Е.А. Боброву за предварительную подготовку исходных данных и проведение расчетов.

Литература

[1] Andrieu, C., de Freitas, N., Doucet, A., Jordan, M.: An Introduction to MCMC for Machine Learning. Machine Learning, 50, pp. 5-43 (2003)

[2] Holten, D.: Hierarchical Edge Bundles: Visualization of Adjacency Relations in Hierarchical Data. IEEE Transactions on Visualization and Computer Graphics, 12, pp. 741-748 (2006)

[3] Karrer, B., Newman, M.: Stochastic Blockmodels and Community Structure in Networks. Physical Review E, 83 (2011)

[4] Peixoto, T.: Nonparametric Bayesian Inference of the Microcanonical Stochastic Block Model. Physical Review E, 95 (2017)

[5] Web of Science, https://www.webofknowledge.com

Альтернативная модель сходства символьных строк

© С.В. Знаменский[1] © В.А. Дьяченко[2]

[1]Институт программных систем имени А.К. Айламазяна РАН,
Переславль-Залесский, Россия
[2] Ярославский государственный университет имени П.Г. Демидова,
Ярославль, Россия

svz@latex.pereslavl.ru dyachenko.vlad_76@mail.ru

Аннотация. Выразительные примеры показывают, что нормализация меры сходства, равно как и замена её метрикой могут приводить к ошибкам кластеризации и ранжирования по сходству. Для задач, в которых сходство определяется выравниванием, описан аналог OCS длиннейшей общей подпоследовательности LCS (Longest Common Subsequence). Предлагаемая модель отвечает потребностям базовых приложений, в которых совпадение подстрок более значимо, чем совпадение разреженных подпоследовательностей той же длины. OCS обещает ускорить приближённый поиск, но точное вычисление за счёт более тонкой градации значений требует умеренных дополнительных ресурсов по сравнению с LCS. Общие диапазоны значений и базовые свойства упрощают миграцию работающих на LCS приложений.

Ключевые слова: мера сходства, метрика близости, LCS, общая подпоследовательность, выравнивание последовательностей, кластеризация, ранжирование.

An Alternative Model of the Strings Similarity

© Sergej Znamenskii[1] © Vladislav Dyachenko[2]

[1]A.K. Ailamazyan Program Systems Institute of Russian academy of Science,
Pereslavl-Zalessky, Russia
[2] P.G. Demidov Yaroslavl State University,
Yaroslavl, Russia

svz@latex.pereslavl.ru dyachenko.vlad_76@mail.ru

Abstract. Expressive examples show that either the normalization of the similarity measure or its replacement by metrics may lead to errors of clustering and ranking by similarity. For applications of alignment-based similarity, an analog OCS of the longest common subsequence (LCS) is described. A proposed model responds the needs of the basic LCS applications in which the matching of substrings is more significant than the coincidence of sparse subsequences of the same length. OCS promises to speed up fuzzy search, but accurate calculation due to a finer gradation of values requires moderate additional resources compared to LCS. Common value ranges and basic properties make it easy to migrate applications running on LCS.

Keywords: similarity measure, similarity metric, LCS, common subsequence, sequence alignment, clustering, ranking.

1 Метрики близости и меры сходства

Символьные строки над конечным алфавитом используются для компьютерного представления информации различной природы при поиске плагиата, работе с версиями исходного кода программ, распознавании звуков и поиске мелодий, анализе данных биоинформатики, грубой сортировке сырых текстовых историко-географических данных и в других прикладных задачах.

Обработка информации различной природы, представленной символьными строками, базируется на численных оценках сходства, обеспечивающих возможности ранжирования по сходству, кластеризации и нечёткого поиска и подробно описанных в многочисленных публикациях.

Не только численные значения оценки сходства символьных строк, но и основанные на них результаты ранжирования по сходству или кластеризации существенно зависят от непростого выбора способа количественной оценки пары строк в основном среди двух групп:

Метрики близости оценивают расстояние между строками. Для строки и её подстроки это обычно разность длин. Формально удовлетворяют известным аксиомам метрического пространства.

Меры сходства оценивают размер общей информации либо мощность пересечения множеств

Труды XIX Международной конференции «Аналитика и управление данными в областях с интенсивным использованием данных» (DAMDID/ RCDL'2017), Москва, Россия, 10–13 октября 2017 года

признаков. Для строки и её подстроки – это обычно длина подстроки. Неотрицательны и монотонны по включению. Некоторые из них могут считаться мерами в смысле классической теории меры, остальные (включая LCS) формализуются как нечёткие меры Шоке–Суджено на множестве признаков. Часто называются метриками сходства, что порождает не всегда корректную ассоциацию с метрическим пространством.

Классический подход к построению меры сходства символьных строк a и b состоит в выравнивании строк выделением в каждой из них одинаковых подпоследовательностей символов.

Выбор длиннейшей из всех таких возможных подпоследовательностей LCS (Longest Common Subsequence) численно оценивает близость символьных строк длиной $d_l(a,b)$ выделенной подпоследовательности. Расстояние Левенштейна $d_l(a,b)$ равно количеству символов, не вошедших в длиннейшую общую подпоследовательность. Поэтому $\mu_l(a,b) + d_l(a,b) = |a| + |b|$, где a и b – длины строк.

Вопрос «Сходство или расстояние: Важно ли?» не случайно возник в [6]. С ним связаны распространенные в научной литературе опасные заблуждения.

1.1 Ошибочное использование метрики

Хотя некорректность применения метрики Левенштейна к строкам различной длины замечена ещё в [10], возможная неэквивалентность меры сходства метрике близости отмечена лишь в [8], а возможная несводимость меры сходства к метрике близости – в [4], но уже в [7] желание использовать методы метрического пространства снова провоцирует спорный вывод: «we have explored the relation between the concepts of distance and similarity and shown that adopting the axiomatic definition of similarity as presented here, leads to a spatial interpretation of similarity as “direction”, complementary to distance».

Анализ неудачной попытки нормализации данных НСКФ-2016 выявил плохую работу метрики Левенштейна и контрпример к этому утверждению:

Хотя сходство строки p=«Переславль» со строкой p_z=«Переславль-Залесский», очевидно, значительно выше, чем со строкой t=«Тверь»

$$\mu_l(p,p_z) = 9 > 3 = \mu_l(p,t),$$

но по расстоянию «Переславль» значительно ближе к «Тверь»

$$d_l(p,p_z) = 11 > 7 = d_l(p,t).$$

Мы видим, что упомянутая простая связь НЕ означает, что расстояние и сходство всегда противоположно направлены. Поэтому любой корректный алгоритм, основанный на метрике, ошибётся в этой ситуации. Использование метрики близости в качестве меры сходства на таких данных порождает ошибки в теоретических выводах и приложениях.

Вывод 1. *Метрику близости рискованно использовать для кластеризации (или ранжирования) по сходству строк существенно различной длины:* различия в длинах маскируют сходство.

1.2 Ошибочное нормирование сходства

Хорошо известно, что метрику можно нормировать без потерь. Простая связь меры близости с метрикой сходства скрывает фундаментальные различия. Ошибки, связанные с нормированием, отчётливо видны на строках u = «USA», r = «RUSSIA» и r_f = «RUSSIAN FEDERATION» и мере сходства LCS: ненормированный LCS вполне обоснованно позиционирует «RUSSIA» в два раза ближе к «RUSSIAN FEDERATION», чем к «USA». Однако нормирование по средней длине резко разворачивает неравенство в обратную сторону:

$$\frac{2\mu(u,r)}{|u|+|r|} = \frac{2}{3} > \frac{1}{2} = \frac{2\mu(r,r_f)}{|r|+|r_f|}.$$

Нормализация по [6, 7] даёт ту же ошибку:

$$\frac{\mu(u,r)}{|u|+|r|-\mu(u,r)} = \frac{1}{2} > \frac{1}{3} = \frac{\mu(r,r_f)}{|r|+|r_f|-\mu(r,r_f)}.$$

Менее опасно, но также не корректно в этом примере нормирование к минимальной длине:

$$\frac{\mu(u,r)}{\min(|u|,|r|)} = 1 = \frac{\mu(r,r_f)}{\min(|r|,|r_f|)}.$$

Вывод 2. *Нормирование меры сходства рискованно при значимых различиях в длине строк*

Примерно половина наиболее активно цитируемых и используемых определений мер сходства постулирует диапазон значений $\mu(x,y) \le \mu(x,x) = 1$. Мы видели, как это порождает ошибки при работе со строками.

Сформулированные замечания значимы не только для географических названий, но и для любых приложений, в которых близость длин не является доминирующим признаком сходства. Во всех приложениях, перечисленных в начале статьи, это именно так, и в каждом из них нетрудно привести аналогичные примеры. Исключение, к которому предостережения данной статьи отношения не имеют, – это задача исправления ошибок набора текста с естественным доминированием близости длин.

2 Информативность общей подпоследовательности

Выравнивание строк выделяет общую подпоследовательность. Например, общая подпоследовательность («с», «в») строк «Переславль» и «Москва» соответствует нескольким практически равноценным выравниваниям

Пере--**с**--лавль и --Перес--лавль
----Москв-**а**--- Мо----**с**кв-**а**---

В известных приложениях носителями информации являются подстроки сопоставляемых строк $x = (x_1,\dots,x_m)$ и $y = (y_1,\dots,y_n)$, совпадение

которых $x_{i+s_x} = y_{i+s_y}\ \forall i = 1,\ldots,k$ является *признаком сходства* строк (здесь s_x и s_y – начальные позиции подстрок, а k – их равная длина, причём $(i+s_i, i+s_y)$ – элемент фиксированной общей подпоследовательности). Разность начальных позиций $s_x - s_y$ будем называть смещением общей подстроки. Для строк «RUSSIA» и «USA» таких подстрок в любой общей подпоследовательности не более, чем четыре (U,US,S,A).

Определение 1. *Наиболее значимой общей подпоследовательностью* назовём такую общую подпоследовательность, в которой количество общих подстрок максимально. Это количество названо в [13, 15] мерой сходства NCS и будет обозначаться $\mu_N(x,y)$.

С другой стороны, каждый такой *признак сходства* несёт долю общей информацию. Размер её формально оценить невозможно. В текстах могут совпасть гениальная фраза либо бессмысленный обрывок, но компьютер этого не разберёт. Для простоты удобно считать все их потенциально информационно равноценными. Поскольку каждая строка несёт в себе информацию каждой своей подстроки, то полный объём общей информации – это снова количество всех подстрок $\mu_N(x,y)$.

Общее количество подстрок наглядно показано количеством указывающих на концы подстроки уголков в примерах:

Мера сходства $\mu_N(x,y)$, очевидно, удовлетворяет классическим аксиомам сходства [5, 7]: *неотрицательность*

$$\mu(x,x) \geq 0, \tag{1}$$

симметричность

$$\mu(x,y) = \mu(y,x), \tag{2}$$

самосходство

$$\mu(x,y) \leq \mu(x,x), \tag{3}$$

супераддитивность меры множества общих признаков

$$\mu(x,y) + \mu(y,z) \leq \mu(x,z) + \mu(y,y), \tag{4}$$

также известная как *неравенство покрытия* или *аналог неравенства треугольника*, и, наконец, *индикация совпадения*

$$\mu(x,y) - \mu(y,y) - \mu(x,y) \Leftrightarrow x - y. \tag{5}$$

Кроме классических аксиом, обе меры сходства (NCS и LCS) обладают свойством *монотонности*

$$y \subset z \Longrightarrow \mu(x,y) \leq \mu(x,z), \tag{6}$$

связанным с вложением подстроки в строку, из которой следует $y \subset x \Leftrightarrow \mu(x,y) = \mu(y,y)$, но не вытекает полезное свойство *индикация подстроки*

$$y \subset x \Longrightarrow \mu(x,y) = \mu(y,y), \tag{7}$$

которым обладает NCS, но LCS не обладает. С другой стороны, LCS обладает простой *связью с длиной строки*

$$\mu(x,x) = |x|, \tag{8}$$

но для NCS это неверно. Диапазон значений у NCS шире, чем у LCS: $0 \leq \mu_N \leq \psi(\min\{m,n\})$. Его верхняя граница $\psi(n) = n(n+1)/2$ – *треугольное число*, дающее согласно [14] совокупное количество подстрок строки длины n.

3 Наивный алгоритм вычисления

Предложенный в [14] алгоритм основан на стандартном применении динамического программирования, реализован на C и доступен на CPAN в виде компилируемого подгружаемого модуля для Perl Algorithm::NCS.

Алгоритм имеет очевидную оценку сложности по памяти $O(mn)$ и по времени $O(m^2 n)$ через длины m и n сравниваемых строк.

Алгоритм 1

```
#include <stdlib.h>
#include <string.h>
int t_ocs(char *x, char*y){
 int *d, k, i, j, n, m, diag, t;
 n = strlen(x)+1;
 m = strlen(y)+1;
 diag = n*m+m;
 d=calloc(sizeof(int), diag+n+1);
 for (i=1; i<n; i++){
     diag++;
     for (j=1; j<m; j++){
         d[j*n+i] = d[(j-1)*n+i] > d[j*n+i-
1]
              ?  d[(j-1)*n+i]
              :  d[j*n+i-1];
         if (x[i-1] == y[j-1]){
             d[diag -j]++;
             if  (d[j*n+i]  <  d[j*n+i-n-1]+
d[diag-j])
                 d[j*n+i]  =  d[j*n+i-n-1]  +
d[diag-j];}
             else { d[diag -j] = 0;}}}
 t = d[n*m-1];
 free(d);
 return t;}
```

Для строк небольшой длины алгоритм обладает сходным со стандартным для LCS быстродействием (численный эксперимент описан ниже).

Известные алгоритмы, включая квадратичный [9], требуют квадратичной памяти, что делает их неприменимыми для длинных строк.

Простейший подход к оптимизации LCS по использованию памяти может быть использован и для ускорения работы NCS.

4 Линейный по памяти алгоритм

Назовём *общим окончанием строк* максимальную общую подстроку, содержащую пару последних элементов, и *эффектным окончанием* максимальную часть общего окончания, входящую в некоторую наиболее значимую подпоследовательность. *Стыком общих подстрок* будем называть такое их расположение, при котором между ними нет просвета в одной из сравниваемых строк.

Алгоритм базируется на двух простых леммах, приводимых без доказательства.

Лемма 1. *Если наиболее значимая общая подпоследовательность содержит стык двух общих подстрок с разными смещениями, то продолжаться через стык может только более короткая из них. В случае равных длин ни одна из строк не может быть продолжена через стык.*

Лемма 2. *Эффектное окончание стыкуется в наиболее значимой подпоследовательности с концом максимальной общей подстроки, представленной в подпоследовательности более длинной, чем это окончание, частью.*

Алгоритм использует вспомогательные массивы данных для компактного хранения информации:

`mp[n]` хранит ранее вычисленные значения NCS для пар начальных подстрок;

`mu[n]` сохраняет текущие вычисляемые значения NCS для пар начальных подстрок.

Используются также массивы данных, индекс которых $s = n - j + i$ связан с фиксированной диагональю:

`ls[s]` содержит начальные позиции общих окончаний в x для разных смещений;

`le[s]` содержит начальные позиции *эффектных* (т .е. включённых в наиболее значимую общую подпоследовательность) общих окончаний.

Нулевое значение элемента ls[s] означает несовпадение окончаний и неактуальность соответствующих значений le[s] и me[s].

Введённые массивы занимают $3m + 8n + 5$ ячеек для хранения целых чисел и инициализируются нулями при вызове функции.

Алгоритм 2

```
for ( i=1; i < m+1; i++ ){
for ( j=1; j < n+1; j++ ){
s = j-i+n;
mx = max( mp[j], mu[j-1] );
if ( x[i-1] == y[j-1] ){
if ( ps[s] == 0 ){/* окончание длины 1 */
ps[s] = i;
me[s] = mp[j-1]+1;
if ( mp[j-1] + 1 > mx ){/* эффектное */
pe[s] = i;
mu[j] = mp[j-1] + 1; }
else { /* неэффектное */
pe[s] = i+1;
mu[j] = mx; }}
else { /* длина больше 1 */
me[s] += i + 1- ps[s];
mt=mp[j-1] + i - pe[s] +1;
if ( (me[s] >= mx) && (me[s] >= mt) ){
mu[j] = me[s];
pe[s] = ps[s]; }
else { /* доминирует не me[s] */
if ( (mt >= mx) && (mt >= me[s]) ){
mu[j] = mt; }
else{ /* доминирует mx */
pe[s] = i+1;
mu[j] = mx; }}}}
else{ /* последние символы различаются */
ps[s] = 0;
mu[j] = mx; }}
swap (mu,mp); }
return mp[n];
```

Здесь swap – обмен указателями на массивы. Корректность кода подтверждена вычислением 1000000 случайных строк с алфавитом из 10 цифр в диапазоне длин до 31×110.

5 Общность порядка

Чтобы исправить диапазон значений и масштаб NCS, рассмотрим обратную к $t = \psi(n)$ монотонную вогнутую функцию

$$n = \varphi(t) = \frac{\sqrt{8t+1}-1}{2}.$$

Определение 2. Назовём сходством *общности порядка* (OCS) меру сходства символьных строк, определённую формулой

$$\mu_O\,(x,y)\;=\;\phi(\mu_N\,(x,y)). \qquad (9)$$

Пример 1. Для рассмотренных в начале статьи строк $\mu_O(r,u)\;\approx\;2.37\;<\;3$.

Дробное значение отражает квадратично лучшее разрешение, ценность которого отмечена в [11]. В данном случае оно естественно отражает наличие просвета в выравнивании с «USA», и тем самым «RUSSIA» оказывается уже не в два, а 2.53 раза ближе к «RUSSIAN FEDERATION», чем к «USA».

Теорема 1. *Сходство упорядоченной общности, определённое формулой (9), удовлетворяет всем аксиомам (1)–(8).*

Доказательство. Все аксиомы, кроме (4), несложно вытекают из определения. Аксиома (4) вытекает из следующей леммы.

Лемма 3. *Для конечного объединения непересекающихся отрезков прямой* $U = \cup_{k=1}^{n}[a_k, b_k]$ *положим* $\mu(U) = \sum_{k=1}^{n}\psi(b_k - a_k)$. *Пусть два подмножества* A *и* B *единичного сегмента* $[0.m]$ *имеют границы, состоящие из конечного числа точек. Тогда* $\mu(A) + \mu(B) \leq \mu(A \cap B) + m$.

Доказательство леммы основано на возможности таких перестроек множеств A и B, при которых пары сегментов сливаются в один с сохранением веса μ так, что неравенство леммы усиливается.

6 Производительность алгоритмов

Для сравнения по производительности LCS и NCS/OCS были испытаны классический алгоритм динамического программирования для LCS и вышеприведённые алгоритмы, которые мы обозначим по порядку NCS1 и NCS2.

В цикле длительностью около 20 секунд генерировались две строки заданных длин, случайно (с равной вероятностью и независимо) заполненные буквами из алфавита фиксированного размера (2 или 128), и измерялось сходство между ними. По времени и количеству вычислений определялось среднее время.

Полная серия экспериментов для различных пар длин и мер сходства была повторена три раза и для каждой измеренной величины на одном

Таблица 1: Отношения среднего времени вычисления в наносекундах к произведению длин

Длины строк	2 буквы			4 буквы			16 букв			128 букв		
	LCS	NCS1	NCS2	LCS	NCS1	NCS2	LCS	NCS1	NCS2	LCS	NCS1	NCS2
10 × 10	46.4	54.8	48.4	47.1	52.2	48.5	43.4	49.0	44.1	42.5	45.9	43.3
10 × 80	12.4	21.4	15.1	12.8	18.5	14.0	11.3	15.4	11.6	10.3	14.1	10.6
80 × 10	12.2	22.0	15.1	12.2	18.1	13.5	11.2	14.9	11.7	10.4	13.9	10.7
10 × 320	7.8	15.1	9.6	8.1	13.3	9.2	7.4	11.6	7.6	6.8	10.4	7.0
320 × 10	7.8	17.7	11.3	7.5	13.7	9.7	7.2	10.9	7.4	6.8	10.2	6.9
100 × 100	9.2	18.4	10.4	8.2	13.3	8.6	6.5	10.2	6.4	6.0	9.3	5.8
100 × 800	4.4	15.0	6.6	4.4	10.5	5.7	4.1	7.8	3.9	3.7	7.0	3.5
800 × 100	4.9	15.1	8.2	4.8	10.6	6.4	4.4	7.6	4.0	3.9	7.0	3.4
100 × 3200	4.0	14.4	5.2	3.8	10.2	4.6	3.8	7.5	3.6	3.6	6.8	3.2
3200 × 100	6.8	14.5	7.9	5.6	10.4	6.0	5.0	8.0	3.6	4.8	7.4	3.2
1000 × 1000	6.9	15.6	8.5	6.0	11.2	6.2	4.7	7.9	3.7	4.1	7.1	3.1
1000 × 8000	8.8	16.0	7.6	9.8	12.1	5.8	9.1	9.7	3.4	8.7	8.7	2.8
8000 × 1000	8.7	17.3	8.2	7.4	12.6	5.9	6.8	10.3	3.4	6.1	9.9	2.8
1000 × 32000	13.8	19.2	7.0	12.6	15.3	5.6	11.7	12.5	3.3	11.7	12.0	2.8
32000 × 1000	20.5	36.0	7.8	19.3	30.4	5.8	18.4	27.4	3.3	18.1	26.4	2.8

персональном компьютере Linux PC Intel(R) Core(TM) i3-3250 CPU @3.50GHz 4 ядра (27935.67 BogoMIPS) 8Gb RAM. Для компенсации случайных флуктуаций были отброшены максимальное и минимальное из каждой тройки полученных значений. Результаты представлены в Таблице 1.

Верхняя часть таблицы показывает, что накладные расходы вызова процедуры доминируют примерно до $nm \approx 10000$, а при большем произведении длин вступают в силу особенности алгоритмов. Первый алгоритм NCS оказывается в 2–3 раза медленнее, чем LCS, а второй примерно на 30% медленнее при малых длинах, но неожиданно быстрее LCS в 2–4 раза на больших.

Возможная причина в том, что даже при формально большем числе операций массивы компактного хранения могут реже требовать изменений (если данное не изменилось, запись не производится), а операции чтения и особенно сравнения выполняются быстрее, чем операции записи. Для проверки этой гипотезы в алгоритме строки `mu[j] = mx;` были заменены на

```
if (mu[j] != mx)
 mu[j] = mx;
```

и аналогично дополнена строка `ps[s] = 0;` в результате время практически не изменилось для бинарного алфавита, но однозначно сократилось в среднем примерно на 0.15 нс для алфавита из 128 символов, что подтвердило гипотезу. Другая возможная причина в том, что процедуры, работающие с данными небольших размеров, могут исполняться в кэше процессора с более быстрыми обращениями к памяти. В любом случае результаты тестов убедительно показали, что на этих случайных данных скорость работы алгоритмов достаточно близка, и разумно организовать эксперимент на реальных данных.

Важно отметить, что все известные оптимизации LCS теряют преимущества при работе со случайными бинарными последовательностями, и цифры в левой половине таблицы и верхних строках, по-видимому, не улучшаемы для алгоритмов, работающих с разными алфавитами. Для ситуации редких совпадений, представленной правой половиной таблички и длинных строк, важной для многих прикладных областей, хорошо известен ряд алгоритмов вычисления LCS, которые останутся вне конкуренции как минимум до тех пор, пока не проработана аналогичная оптимизация для NCS.

7 Путь к быстрым приближённым алгоритмам

Поскольку квадратичная сложность неприемлема для поиска сходных подстрок в большой базе экспериментальных данных, то острой является потребность в быстрых алгоритмах, надёжно отсеивающих основную часть несхожей информации, чтобы малую оставшуюся часть обработать алгоритмами, точно оценивающими сходство.

Корень проблемы в том, что любой алгоритм оценки сходства символьных строк базируется на попарном сравнении элементов.

Гипотеза 1. *Пусть в квадратной таблице размером* $2n \times 2n$ *отмечено* n^2 *клеток. Тогда в ней можно выбрать последовательность из* n *неотмеченных клеток, у которой номера строк и номера столбцов строго возрастают.*

Отметка клеток, наиболее удалённых от побочной диагонали, образующая два треугольника, один чуть больше другого, по-видимому даёт ситуацию, единственную с точностью до центральной симметрии, в которой более длинных последовательностей нет.

Если гипотеза 1 верна, то LCS принципиально не допускает приближённых алгоритмов поиска с лучшей, чем квадратичная, оценкой, пригодных для предварительной фильтрации при поиске. Это согласуется с давно известной [2] невозможностью точного работающего с алфавитами любых размеров алгоритма для LCS с лучшей, чем квадратичная, оценкой сложности.

Благодаря чувствительности к просветам, ситуация для OCS (и NCS) отличается принципиально:

Теорема 2. *Для любого* $\varepsilon > 0$ *существует такой*

номер $N > 0$, *что при любых* $m, n > N$ *можно указать менее* $m\varepsilon^{-2}$ *пар элементов, несовпадение которых влечёт неравенство* $\mu_O(x, y) < \varepsilon\, n$.

Сформулированная теорема по сути означает существование алгоритма предварительной фильтрации при поиске, использующего сравнение $m\,\varepsilon^{-2}$ пар элементов. При фиксированной относительной погрешности ε это означает линейную сложность алгоритма для $m = n$. Искомый алгоритм может быть получен предположительно несложной доработкой NCS2.

Доказательство. Зафиксируем k, равное целой части от $(n+1)\varepsilon^2$, и рассмотрим множество из не более чем $\frac{mn}{k}$ пар: $\{(i, j): j \bmod k = 0\}$. Любая общая подпоследовательность вне этого множества пар не может иметь вес, больший чем

$$\phi\left(\left(\frac{n}{k}+1\right)\psi(k-1)\right) = \phi\left(n\frac{k-1}{2}\right) \le \varepsilon n. \square$$

Можно предположить, что десятилетия интенсивных многоплановых поисков [1] замены базовых эвристических алгоритмов биоинформатики не уступающими в производительности, но аккуратно теоретически обоснованными, не дали результата потому, что поиски велись вдали от OCS.

8 Устойчивость к случайному шуму

Канонический набор данных для тестирования мер сходства составляют случайно генерированные строки, позволяющие объективно оценить важные для приложений качества мер сходства. В частности, строки четырёхбуквенного алфавита с независимым и равномерным распределением букв успешно моделируют объекты биоинформатики.

В ходе численного эксперимента на том же компьютере получены представленные на Рис. 1 кумулятивные гистограммы (эмпирические функции распределения) значений мер сходства LCS и OCS строк фиксированных длин из равновероятных независимых букв четырёхбуквенного алфавита.

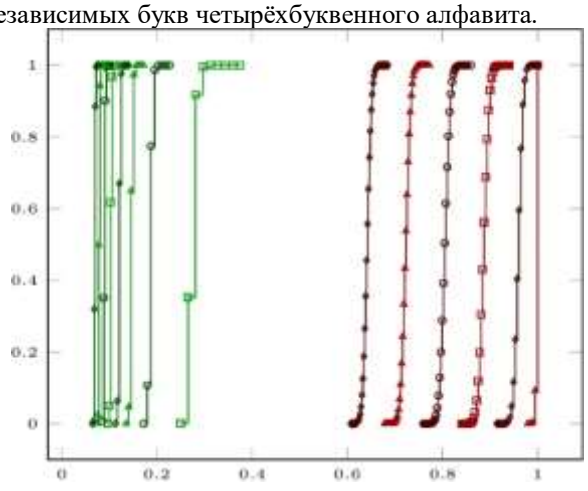


Рисунок 1 Кумулятивные гистограммы меры общности пары случайных строк суммарной длины $m + n = 1024$; слева направо отношения длин последовательностей m:n= 1:1, 7:9, 3:5, 5:11, 1:3, 3:13, 1:6, 1:15; семейство графиков для $\frac{\mu_O(x,y)}{m}$ расположено левее семейства для $\frac{\mu_L(x,y)}{m}$, два крайних справа представлены точкой (1,1)

При отношении длин 3:13 и ниже для строк суммарной длины 1024 символа мы не получаем из LCS никакой информации, поскольку результат предопределён с вероятностью, близкой к 1. При рассмотренных отношениях строк LCS, как правило, превышает 0.6 от максимального значения, а диапазон изменения NCS оказывается больше в разы. Низкие математическое ожидание и дисперсия означают низкую вероятность значимого сходства, которую можно интерпретировать как устойчивость к случайному шуму.

Рис. 2 показывает, что при увеличении алфавита ситуация меняется количественно, но качественный разрыв сохраняется.

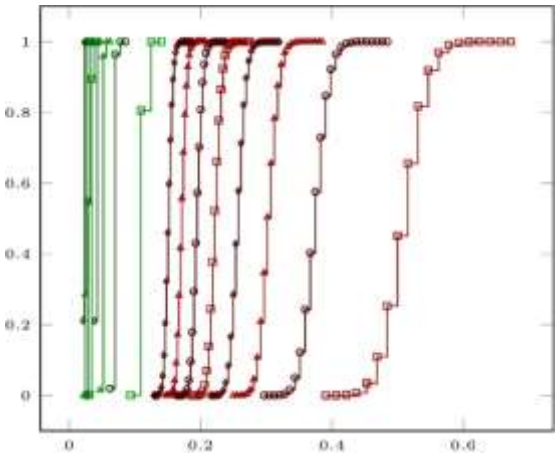


Рисунок 2 Кумулятивные гистограммы меры общности пары случайных строк суммарной длины $m + n = 1024$ для алфавита из 128 букв; слева направо отношения длин последовательностей m:n= 1:1, 7:9, 3:5, 5:11, 1:3, 3:13, 1:6, 1:15; семейство графиков для $\frac{\mu_O(x,y)}{m}$ расположено левее семейства для $\frac{\mu_L(x,y)}{m}$

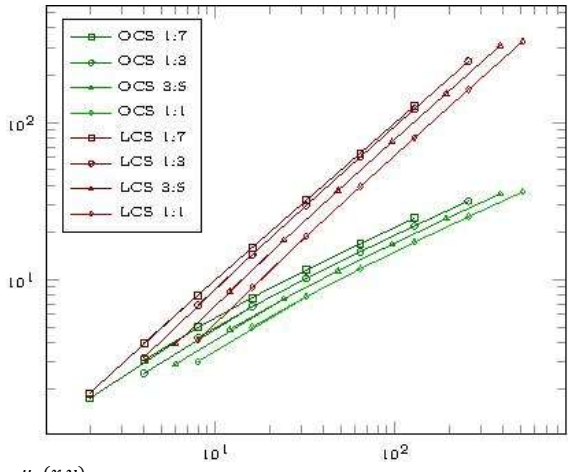


Рисунок 3 Зависимость математического ожидания мер общности для четырёхбуквенного алфавита от длины кратчайшей из строк при фиксированных отношениях длин

Рис. 3 и 4 демонстрируют значимое усиление эффекта по мере роста длин сравниваемых строк.

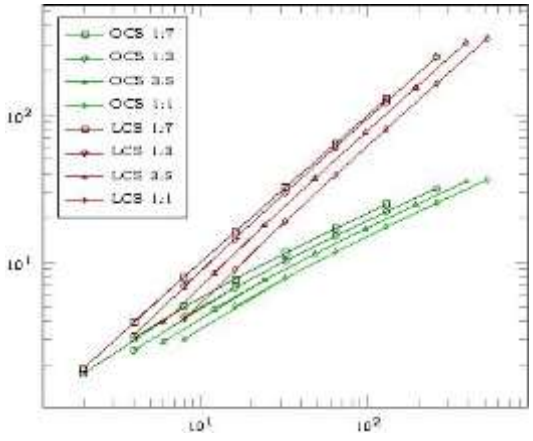


Рисунок 4 Зависимость дисперсии мер общности для четырёхбуквенного алфавита от длины кратчайшей из строк при фиксированных отношениях длин

При отношении длин 3:13 и ниже для строк суммарной длины 1024 символа мы не получаем из LCS никакой информации, поскольку результат предопределён с вероятностью, близкой к 1. При рассмотренных отношениях строк LCS, как правило, превышает 0.6 от максимального значения, а диапазон изменения NCS оказывается больше в разы. Низкие математическое ожидание и дисперсия означают низкую вероятность значимого сходства, которую можно интерпретировать как устойчивость к случайному шуму.

Заключение

Описана мера близости, обладающая естественным определением, уникальным сочетанием полезных аксиом (1)–(8), сопоставимыми по сложности и скорости алгоритмами, повышенными разрешением и устойчивостью к случайному шуму в сравнении с LCS.

Выбор других мер сходства для объективного сравнения сильно затруднён множественностью потенциально возможных интуитивно непрозрачных настроек, таких, как коэффициенты широко используемой в биоинформатике функции просветов (gap function) [3, 12].

Литература

[1] Abboud, A., Williams, V.V., Weimann, O.: Consequences of Faster Alignment of Sequences. Int. Colloquium on Automata, Languages, and Programming. Springer Berlin Heidelberg, pp. 39-51 (2014)

[2] Aho, A.D., Hirschberg, D.S., Ullman, J.D.: Bounds on the Complexity of-the Maximal Common Subsequence Problem. JACM, 23 (1), pp. 1-12 (1976)

[3] Cartwright, R.A.: Logarithmic Gap Costs Decrease Alignment Accuracy. BMC Bioinformatics, 7, 527 (2006)

[4] Chen, M., Li X., Ma, B., Vitányi, P.M.: The Similarity Metric. IEEE Transactions on Information Theory, 50 (12), pp. 3250-3264 (2004)

[5] Chen, S., Ma, B., Zhang, K.: On the Similarity Metric and the Distance Metric. Theoretical Computer Science, 410 (24–25), pp. 2365-2376 (2009)

[6] Elzinga, C.H.: Distance, Similarity and Sequence Comparison. Advances in Sequence Analysis: Theory, Method, Applications. Springer International Publishing, pp. 51-73 (2014)

[7] Elzinga, C.H., Studer, M.: Normalization of Distance and Similarity in Sequence Analysis. LaCOSA II, Lausanne, June 8–10, pp. 445-468 (2016)

[8] Emms, M., Franco-Penya, H.H.: On the Expressivity of Alignment-Based Distance and Similarity Measures on Sequences and Trees in Inducing Orderings. Springer Proceedings in Mathematics & Statistics, 30, pp. 1-18 (2013)

[9] Guo, Y.-P., Peng, Y.-H., Yang, C.-B.: Efficient Algorithms for the Flexible Longest Common Subsequence Problem. Proc. of the 31st Workshop on Combinatorial Mathematics and Computation Theory, pp. 1-8 (2014)

[10] Lim, S. Cleansing Noisy City Names in Spatial Data Mining. 2010 Int. Conf. on Information Science and Applications (ICISA), p. 18 (2010)

[11] Tseng, K.-T., Yang, C.-B., Huang, K.-S.: The Better Alignment Among Output Alignments. J. of Computers, 3, pp. 51-62 (2007)

[12] Wang, C., Yan, R.X., Wang, X.F., Si, J.N., Zhang, Z.: Comparison of Linear Gap Penalties and Profile-Based Variable Gap Penalties in Profile-Profile Alignments. Computational Biology and Chemistry, 35 (5), pp. 308-318 (2011)

[13] Znamenskij, S.V.: A Model and Algorithm for Sequence Alignment. Program systems: theory and applications, 6 (1), pp. 189-197 (2015)

[14] Znamenskij, S.V.: Simple Essential Improvements to ROUGE-W algorithm. J. of Siberian Federal University. Mathematics & Physics, 4, pp. 258-270 (2015)

[15] Znamenskij, S.V.: A Belief Framework for Similarity Evaluation of Textual or Structured Data. Similarity Search and Applications, LNCS 9371, pp. 138-149 (2015)

Онтологические модели и применения 1

Ontological models and applications 1

A Domain-Agnostic Tool for Scalable Ontology Population and Enrichment from Diverse Linked Data Sources

© Efstratios Kontopoulos © Panagiotis Mitzias © Marina Riga © Ioannis Kompatsiaris

Information Technologies Institute,
GR-57001 Thessaloniki, Greece

skontopo@iti.gr pmitzias@iti.gr mriga@iti.gr ikom@iti.gr

Abstract. Ontologies are a rapidly emerging paradigm for knowledge representation, with a growing number of applications in data-intensive domains. However, populating enterprise-level ontologies with massive volumes of data is a non-trivial and laborious task. Towards tackling this problem, the field of ontology population offers a multitude of approaches for populating ontologies with instances in an automated or semi-automated way. Nevertheless, most of the related tools typically analyse natural language text and neglect more structured types of information like Linked Data. The paper argues that the rapidly increasing array of published Linked Datasets can serve as the input for large-scale ontology population in data-intensive domains and presents PROPheT, a novel software tool for ontology population and enrichment. PROPheT can populate a local ontology model with instances retrieved from diverse Linked Data sources served by SPARQL endpoints. As demonstrated in the paper, the tool is domain-agnostic and can efficiently handle vast volumes of input data. To the best of our knowledge, no existing tool can offer PROPheT's diverse extent of functionality.

1 Introduction

Ontologies constitute a knowledge representation paradigm for modelling domains, concepts and interrelations, effectively enabling the sharing of information between diverse systems [23].The rapidly emerging popularity of ontologies has led to their deployment in various *Data Intensive Domains* (*DID*s), like e. g. bioinformatics [7], e-commerce [11] and digital libraries [3]. Nevertheless, in order for ontologies to be further used at an enterprise level, massive volumes of data are required for populating the underlying models.

If performed manually, this task is extremely time-consuming and error-prone. *Ontology population* attempts to alleviate this problem, by introducing methods and tools for automatically augmenting an ontology with instances of concepts and properties. The schema of the ontology itself is not altered but only its set of concepts and relations. This process is part of *ontology learning*, which refers to the automatic (or semi-automatic) construction, enrichment and adaptation of ontologies [16].

The vast majority of ontology population tools and methodologies are aimed at textual input, typically extracting knowledge from natural language text [5], [20]. However, other more structured sources of information are very often neglected. Such an example is *Linked Data* [10], which builds upon standard Web technologies and is a standard for publishing interlinked structured data that are capable of responding to semantic queries. Linked Data are formalised using controlled vocabulary terms based on ontologies and can be publicly accessible via a SPARQL endpoint [4].

This paper argues that the rapidly increasing array of published Linked Datasets [1] can serve as the input for large-scale ontology population in DIDs and presents PROPheT, a software tool for user-driven ontology population from Linked Data sources. The tool is domain-agnostic and can efficiently handle vast volumes of input data. To the best of our knowledge, no existing tool can offer PROPheT's extent of functionality.

The rest of the paper is structured as follows: Section 2 gives an overview of related work approaches. Section 3 presents PROPheT in detail, followed by a discussion on PROPheT's performance with regards to key challenges for accessing information served by SPARQL endpoints. Section 5 presents an illustrative use case that demonstrates the tool's versatility and scalability. Section 6 presents an evaluation of PROPheT, and the paper is concluded with final remarks and directions for future work.

2 Related Work

Ontology population has already been deployed in various domains, like e.g. e-tourism [22], web services [21] and clinical data [17], amongst others. Regarding the application of ontology population in DIDs, we only came across a recent (2016) work by Knoell et al. revolving around Big Data [15], indicating a potentially emerging interest in the area.

Overall, and as already mentioned in the introduction, state-of-the-art ontology population approaches in literature are mostly addressed to retrieving instances from textual corpora (i.e. natural language text, like e.g.

Proceedings of the XIX International Conference "Data Analytics and Management in Data Intensive Domains" (DAMDID/RCDL'2017), Moscow, Russia, October 10–13, 2017

product catalogues) and mainly involve machine learning, text mining and natural language processing (NLP) techniques. Other indicative approaches besides the ones discussed above are presented in [5] and [20].

A less popular stream of ontology population research is aimed at retrieving instances from other types of content, like e.g. CAD files [8], or more structured content, like e.g. spreadsheets [9], [13], and XML files [19]. However, to the best of our knowledge, no other approach similar to PROPheT currently exists that is capable of populating an ontology with instances retrieved from Linked Data sources, rendering PROPheT into a highly novel tool.

3 The PROPheT Ontology Population Tool

PROPheT[1] is a novel software tool for ontology population and enrichment that can retrieve instantiations of concepts from SPARQL-served Linked Data sources in a scalable manner. The retrieved instances are filtered based on user preferences and are then inserted into a target ontology. As described in the following subsections, PROPheT provides various modes of instance retrieval, along with the capability for establishing user-defined mappings of the respective properties. The tool's mode of operation is purely user-driven, but relies on a step-by-step wizard-based interaction with the end-user, which greatly facilitates use of the software even by largely unfamiliarised users.

PROPheT's front-end (see main window in Figure 1) relies on Python and the PyQt application framework, while the back-end deploys RDFLib and SPARQLWrapper, two Python APIs for manipulating ontologies, along with an SQLite data store for storing settings and user preferences.

PROPheT is fully domain-independent in the sense that it can operate with *any* OWL ontology and *any* RDF Linked Dataset that is served via a SPARQL endpoint.

3.1 Motivation

PROPheT was developed within the recently finished PERICLES FP7 project on Digital Preservation[2]. One of the domains tackled by the project was cultural heritage, where we faced the non-trivial challenge of populating our domain ontologies with thousands of artefacts, each of which was associated with hundreds of metadata entries. In this affair, PROPheT was successfully deployed for populating the ontologies with instances retrieved from various Linked Data sources, like DBpedia and Freebase.

Nevertheless, though highly relevant [26], cultural heritage is not the only DID where populating ontologies from diverse sources poses a formidable challenge. Other domains share similar concerns, like e.g. the telecommunications and news industry [2], and health and biomedicine [6], [14]. This was our main motivation for turning PROPheT into a truly domain-agnostic tool, capable of performing ontology population and enrichment from Linked Data sources in virtually any domain, data-intensive or not.

3.2 Ontology Population

PROPheT offers the capability of class-based and instance-based ontology population. The former method, *class-based population*, retrieves instances from an external model and inserts them into a local ontology, based on a class name entered by the user. Since the exact class name has to be entered (e.g. `dbo:Artist` for the DBpedia class representing artists), this method has the peculiarity that the user needs to know the structure of the external ontology. PROPheT then submits appropriate SPARQL queries to the remote model's endpoint and retrieves a result set of instances belonging to the specified class. The user may then select the instances to populate an existing class in the local ontology.

The second method, *instance-based population*, has two different modes:

(a) Retrieval based on instance label, which is performed according to a label (`rdfs:label` property value) entered by the user. The match of the retrieved instances is based on an exact or partial match of the input text.

(b) Retrieval based on an existing instance, in which the user selects an instance already existing in the local ontology and PROPheT queries the endpoint for similar instances. More specifically, the tool finds classes in the remote ontology that include an instance with a similar `rdfs:label` property value with the input instance. The user may then select specific classes, view their extension (i.e. set of instances) and choose which instances to import into the local ontology model.

In all the cases described above, and after the set of preferred instances has been selected by the user to be populated into the ontology, PROPheTlaunches the ontology mapping process described next.

3.3 Ontology Mapping

In order for PROPheT to proceed with the ontology population with the selected instances, a user-driven ontology mapping is performed, in the sense that the properties of the retrieved instances have to be mapped to properties defined in the local model. In this context, PROPheT displays to the user a list of all datatype properties (`owl:DatatypeProperty`) for the selected instances, in order for him/her to manually define appropriate mappings to datatype properties already existing in the local ontology. This mapping between local and remote properties is mandatory for the property values to be inserted into the local ontology along with the instances. For example, the user might define that the retrieved property `dbo:birthDate` corresponds to the local property `ex:dateOfBirth`. Once defined by the user, PROPheT stores the mappings and offers suggestions when the same mappings occur again

[1]PROPheT is available at: http://mklab.iti.gr/project/prophet-ontology-populator

[2] http://www.pericles-project.eu/

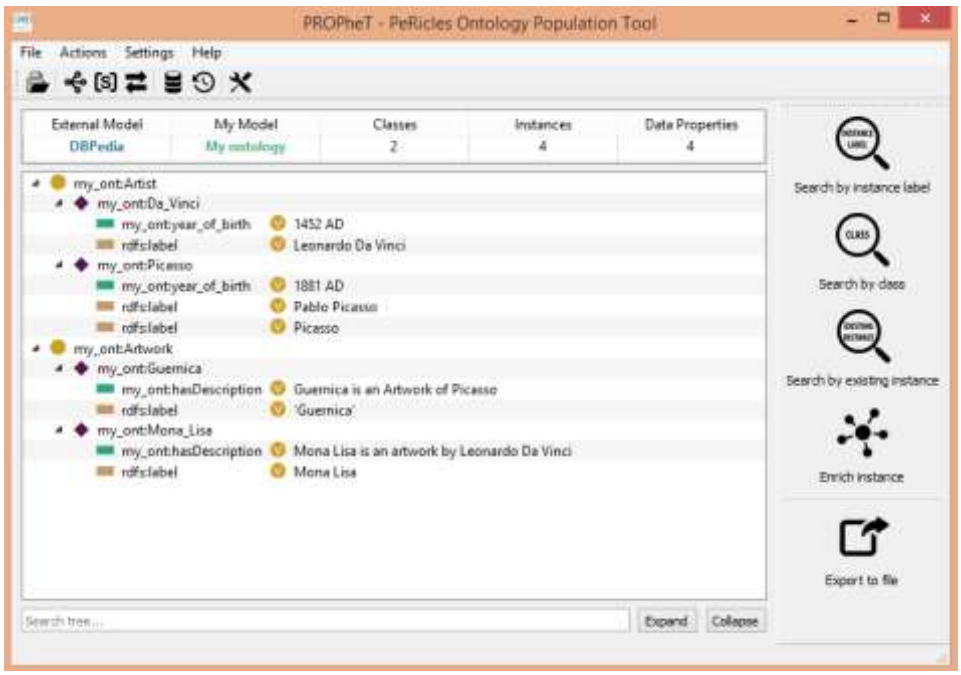


Figure 1 PROPheT's main window

3.4 Instance Enrichment

Besides the ontology population capabilities described above, PROPheT also offers the option of enriching instances already existing in the local ontology with properties and values from "similar" instances in remote ontologies; instance similarity here refers to similarity in the respective instance labels (i.e. `rdfs:label`).

The similar instances may belong to one or more different classes in the remote ontology, thus, the tool presents the user with the type (`rdf:type` property declaration) of each instance. Based on the content and semantics of the derived instances, the user may then decide which property-value pairs he/she will insert from the remote into the local ontology.

3.5 Ontology Enrichment

The local model may also be semantically enriched by establishing links between properties in the local and the remote ontologies via `owl:equivalentProperty` declarations added into the local model. Similar links between classes are represented via `owl:sameAs` and/or `rdfs:seeAlso` declarations added to the local ontology.

4 Challenges and PROPheT's Performance

The availability and scalability of the SPARQL endpoints serving Linked Data is not always guaranteed, since maintaining such heavyweight query services implies significant server-side costs coupled with various potential technical problems on the level of the infrastructure itself [4]. Key parameters for evaluating a SPARQL endpoint are [4]:

- *Discoverability*, referring to how an endpoint can be located and what are the available metadata;
- *Interoperability*, with regards to the supported SPARQL version(s);
- *Efficiency*, which relates to the time needed to respond to the query;
- *Reliability*, based on the uptime of the endpoint on a constant basis.

A useful tool for monitoring the above parameters of SPARQL endpoints is *SPARQLES* [24], while the recent *Linked Data Fragments* (*LDF*) paradigm promises to alleviate the burden from endpoints, by redistributing the load between clients and servers [25].

Taking the above challenges into consideration, and in order to demonstrate PROPheT's scalability, we experimented with timing the retrieval and population of instances from the following well-known SPARQL endpoints into a local custom ontology model:

- *DBpedia*[1], the Linked Data version of WikiPedia;
- *OpenDataCommunities*[2], the official Linked Data platform of the UK Department for Communities and Local Government (DCLG) that provides a selection of official statistics and data outputs on a variety of themes related to DCLG;
- *DBLP*[3], which provides open bibliographic information on major computer science journals and proceedings;
- The *Nobel Prize Linked Data dataset*[4] that contains the authoritative information about Nobel prizes and

[1] http://dbpedia.org/sparql
[2] http://opendatacommunities.org/sparql
[3] http://dblp.l3s.de/d2r/sparql
[4] http://data.linkedmdb.org/sparql

Nobel Laureates since 1901;

- *Eurostat statistics*[5] converted to RDF and re-published using Linked Data principles.

Table 1 illustrates the resulting retrieval and population times for all selected endpoints. PROPheT's performance is impacted by three parameters: (a) the software's efficiency in querying and handling data, (b) the endpoints' speed in serving the requested data, and (c) the volume of data (in the form of datatype property values) that the retrieved instances are attached to.

Table 1 Instance retrieval and population times

Ontology	No of instances	Retrieval time (sec)	Population time (sec)
DBpedia	10	6,0	2,5
	100	19,0	8,3
	1,000	171,0	54,0
	10,000	648,0	250,0
Open Data Communities	10	4,5	3,0
	100	18,0	6,7
	1,000	104,0	44,0
	10,000	510,0	210,0
DBLP	10	3,5	1,8
	100	10,0	5,0
	1,000	62,0	32,0
	10,000	316,0	192,0
Nobel Prize	10	3,7	2,0
	100	10,0	5,7
	1,000	56,0	31,0
	10,000	270,0	170,0
Eurostat	10	5,0	2,5
	100	15,7	7,7
	1,000	92,0	48,0
	10,000	440,0	225,0

Since parameter (a) remains constant within the experiments, it becomes obvious that any differentiation in times heavily depends on parameters (b) and (c). Considering the facts that DBpedia reportedly contains the largest volume of property values, that most of the rest endpoints had almost equal number of properties and that Eurostat's selected instances had no datatype properties, it is clear that an endpoint's response time (second parameter) has a great impact on the ontology population process from Linked Data sources.

5 A Use Case Scenario

This section intends to demonstrate PROPheT's functionality by presenting a use case scenario in a data-intensive domain. Thus, consider a government institution monitoring pollution in rural environments, which requires a directory of cities and towns worldwide, enriched with related information, such as population, postal codes, etc.

Initially, a local ontology schema needs to be deployed, incorporating the necessary classes (e.g. *Town*, *City*, etc.) and properties (e.g. *hasPopulation*, *hasPostalCode*, etc.). This schema will be loaded in PROPheT to be populated.

Next, the user will need to register the sources that serve the desired data (SPARQL endpoint URIs). For the domain of the specific use case, there are several established SPARQL-served ontologies that contain instances of cities and towns, such as *ENVO*[6], an ontology of environmental features and habitats, and *LinkedGeoData*[7]. Specifically, ENVO's class *City* (ENVO_00000856) and LinkedGeoData's classes *City* and *Town* contain related instances.

Table 2 Instance retrieval and population times

Ontology	No of instances	Population time (sec)
LinkedGeoData	10,000	120
ENVO	10,000	204
LinkedGeoData	10,000	158

Taking advantage of PROPheT's class-based instance extraction wizard, the user can respectively populate two (or more) different classes of the local schema with resources from two (or more) data sources. For the purposes of this case study, PROPheT flawlessly managed to retrieve and populate more than 30K instances, along with data property values. Specifically, 10K instances from ENVO's *City* and 10K instances from LinkedGeoData's *City* were populated in the local model's class *City*. Also, another 10K instances from LinkedGeoData's class *Town* were imported to the local model's *Town*. Indicatively, Table 2 displays the population times (in seconds) for the instances mentioned above. Population times in the second batch of LinkedGeoData instances is slightly higher, since the local ontology already contained 20K instances populated during the previous two phases.

Alternatively, supposing that the user cannot predefine the classes of interest in the external models, a different course will be followed. First, a single instance of the desired set will be located and imported. For example, using PROPheT's feature "*Search by instance label*", the user will find a certain city of interest, e.g. Amsterdam, and import it into the local model. Next, with the use of "*Search for similar instances*", the software will discover all the classes where Amsterdam is assigned to. Browsing the resulting list of classes, the user will now locate the classes of interest (e.g. class *City*) and select more instances to be populated.

Consequently, by utilizing the "*Enrich Instance*" function, the user can semantically enrich the major cities' instances (e.g. London, Paris, Amsterdam) with data regarding air pollution levels, residing in different endpoints.

To conclude, the aforementioned use case demonstrates PROPheT's ability to populate various classes of an ontology with data retrieved from more than one endpoints

[5] http://eurostat.linked-statistics.org/sparql
[6] http://www.obofoundry.org/ontology/envo.html
[7] http://linkedgeodata.org

.

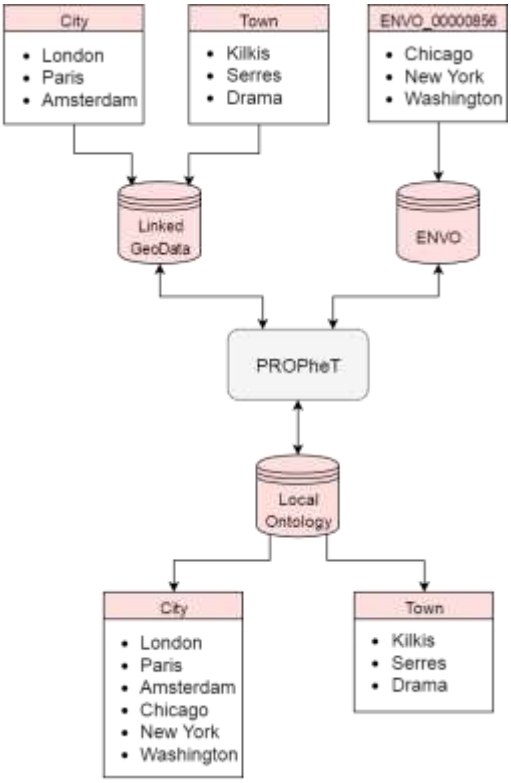


Figure 2 Use case diagrammatic overview

Figure 2 illustrates a diagrammatic overview of the use case described in this section. The population-related features permit different approaches for searching and browsing the available information, offering, thus, great flexibility to the user.

6 PROPheT Evaluation

We recently conducted a user evaluation on PROPheT with very positive results [18]. As indicated by the resulting evaluation of the participants, the following aspects of the tool were the most positive ones: attractiveness (93.5%), user-friendliness (93.5%), ease of usage (100%), innovativeness (87.5%), and efficiency (93.5%); the numbers in parentheses correspond to the respective percentages indicating acceptance on behalf of the users. The current section now presents a qualitative evaluation of the tool, based on the categorisation criteria for ontology population tools proposed in [20].

Elements extracted: Refers to the capacity of an ontology populating system to extract the various ontological aspects, like e.g. objects and relations. PROPheT offers the capability of extracting from external sources both objects (i. e. class instances) and relations (i.e. data property values), and inserts them into a local ontology model. Additionally, PROPheT also appends properties for semantically enriching the local model via `owl:equivalentProperty`, `owl:sameAs` and `rdfs:seeAlso`.

Initial requirements: This criterion refers to the system's initial requirements in terms of resources or background knowledge. PROPheT's only requirement is that a local OWL ontology is already available, in order to be populated with objects retrieved from Linked Data sources. No domain-dependant resources are needed, since PROPheT can flexibly adapt to any thematic domain. No specialised software should be installed in the host machine either; PROPheT is distributed as a standalone bundle.

Learning approach: Refers to the system's approach in extracting knowledge and whether this approach is specialised to a domain. Ontology population tools typically employ Machine Learning techniques (see Section 2), via statistical methods to identify terms or via automated pattern extraction. PROPheT, on the other hand, deploys a purely user-driven, step-by-step ontology population and enrichment approach, which is suitable even for users with only fundamental familiarity with the pertinent notions.

Degree of automation: A fully automated ontology population system is of course desirable, but it seldom is possible to achieve, as the involvement of a domain expert or an ontology engineer is very often needed. The PROPheT approach is mainly user-driven, requiring the involvement of an end user for performing ontology population and enrichment through a step-by-step wizard-based graphical user interface. Thus, although the tool requires user intervention at each step, the process is achieved in a highly user-friendly fashion, as

demonstrated by our recent user evaluation of the tool [18] that indicated very positive feedback on behalf of the users.

Consistency maintenance and redundancy elimination: This criterion refers to the system's capability to maintain the consistency of the ontology, which is highly crucial, and to reduce redundancy, which is not equally vital but can facilitate the process of querying the ontology and can limit its size and complexity. Consistency maintenance in PROPheT is ensured by the integrated specialised APIs for manipulating ontologies and SPARQL queries. On the other hand, the problem of instance redundancy (i.e. two or more instances in the ontology referring to the same real object) is handled by PROPheT in a way that instances with the same name-identifier cannot be populated multiple times in the ontology, i.e. values of populated data properties are linked to one single instance. Moreover, we are currently investigating adding more complex handling mechanisms, such as heuristics or machine learning methods to identify similar resources.

Domain portability: This is an important aspect for all ontology populating systems and refers to their capability to be ported to multiple thematic domains or not. PROPheT is a totally domain-agnostic tool that is able to operate equally successfully in any domain, as long as the external sources are served through SPARQL endpoints. Thus, no domain-specific knowledge is incorporated into the system.

Corpora modality: A system that is able to process various modalities demonstrates its ability to accommodate and exploit diverse knowledge sources. In this context, PROPheT can only process input from Linked Data sources through SPARQL endpoints, but can easily be extended to process third-party ontologies as well, retrieving instances and enriching the local ontology with additional property values found in these models.

7 Conclusions and Future Work

The paper argued that, with the rapidly emerging advent of the use of ontologies in data-intensive domains, the process of ontology population becomes increasingly relevant. Most proposed solutions are typically aimed at analysing natural language text, often overlooking other sources of more structured information, like e.g. Linked Data. In this context, we presented PROPheT, a domain independent software tool for ontology population and enrichment from Linked Data sources. Through wizard-based user-driven processes, the tool facilitates the semi-automatic retrieval of instances and their insertion into a local OWLontology. An advanced mapping process enables the dynamic definition of matching classes and properties between source and target models. PROPheT's rich functionality and versatility cannot be matched by any other ontology population tool found in literature, making PROPheT a truly innovative system for populating and enriching ontologies.

Nevertheless, there are still a few areas of improvement for the tool. In its current implementation, PROPheT is only limited to handling datatype and not object properties; the latter are significantly more complex to tackle. Additionally, the tool cannot currently handle direct or indirect imports of ontologies. A further improvement could be considering additional semantic enrichment associations, like e.g. `skos:narrower` and `skos:broader` from SKOS [12]. And, finally, the process of suggesting similar instances or classes to the user during the population and enrichment steps could be suggested by the tool itself, according to appropriate similarity metrics. We are currently working on a revised version of the software, which will integrate the improvements mentioned above.

Acknowledgements

This research received funding by the European Commission Seventh Framework Programme under Grant Agreement Number FP7-601138 PERICLES. We would also like to thank the anonymous reviewers for their valuable remarks, thanks to which the paper has been significantly improved.

References

[1] Abele, A., McCrae, J.P., Buitelaar, P., Jentzsch, A., Cyganiak, R.: Linking Open Data Cloud Diagram (2017). http://lod-cloud.net/.

[2] Belam, M.: What is the Value of Linked Data to the News Industry? The Guardian (2010, January). https://www.theguardian.com/help/insideguardian/2010/jan/25/news-linked-data-summit

[3] Buckingham Shum, S., Motta, E., Domingue, J.: ScholOnto: An Ontology-based Digital Library Server for Research Documents and Discourse. International J. on Digital Libraries, 3 (3), pp. 237-248 (2000)

[4] Buil-Aranda, C., Hogan, A., Umbrich, J., Vandenbussche, P.Y.: Sparql Web-querying Infrastructure: Ready for Action? Int. Semantic Web Conf., pp. 277-293. Springer (2013)

[5] Buitelaar, P., Cimiano, P.: Ontology Learning and Population: Bridging the Gap Between Text and Knowledge, 167, Ios Press (2008)

[6] Callahan, A., Cruz-Toledo, J., Dumontier, M.: Ontology-based Querying with Bio2RDF's Linked Open Data. J. of Biomedical Semantics, 4 (1), S1 (2013)

[7] Gene Ontology Consortium. The Gene Ontology (GO) Database and Informatics Resource. Nucleic Acids Research, 32 (suppl 1), D258-D261 (2004)

[8] Häfner, P., Häfner, V., Wicaksono, H., Ovtcharova, J. Semi-automated Ontology Population from Building Construction Drawings. KEOD, pp. 379-386 (2013)

[9] Han, L., Finin, T., Parr, C., Sachs, J., Joshi, A.: RDF123: From Spreadsheets to RDF. Proc. of the 7th Int. Semantic Web Conf. pp. 451-466, Springer (2008)

[10] Heath, T., Bizer, C.: Linked Data: Evolving the Web into a Global Data Space. Synthesis Lectures on the Semantic Web: Theory and Technology, 1 (1), pp. 1-136 (2011)

[11] Hepp, M.: Goodrelations: An Ontology for Describing Products and Services Offers on the Web. Int. Conf. on Knowledge Engineering and Knowledge Management, pp. 329-346. Springer (2008)

[12] Isaac, A., Summers, E.: SKOS Simple Knowledge Organization System. Primer, World Wide Web Consortium (W3C) (2009)

[13] Jupp, S., Horridge, M., Iannone, L., Klein, J., Owen, S., Schanstra, J., … Stevens, R.: Populous: A Tool for Building OWL Ontologies from Templates. BMC Bioinformatics, 13 (Suppl 1), S5 (2012). http://doi.org/10.1186/1471-2105-13-S1-S5.

[14] Jupp, S., Malone, J., Bolleman, J., Brandizi, M., Davies, M., Garcia, L., ... Wimalaratne, S.M.: The EBI RDF Platform: Linked Open Data for the Life Sciences. Bioinformatics, 30 (9), pp. 1338-1339 (2014)

[15] Knoell, D., Atzmueller, M., Rieder, C., Scherer, K.P.: BISHOP-Big Data Driven Self-Learning Support for High-performance Ontology Population. LWDA, pp. 157-164 (2016)

[16] Maedche, A., Staab, S.: Ontology Learning from the Semantic Web. IEEE Intelligent Systems, 16 (2), pp. 72-79 (2001)

[17] Mendes, D., Rodrigues, I.P., Baeta, C.F.: Development and Population of an Elaborate Formal Ontology for Clinical Practice Knowledge Representation. KEOD, pp. 286-292 (2013)

[18] Mitzias, P., Riga, M., Kontopoulos, E., Stavropoulos, T.G., Andreadis, S., Meditskos, G., Kompatsiaris, I. User-Driven Ontology Population from Linked Data Sources. Int. Conf. on Knowledge Engineering and the Semantic Web, pp. 31-41. Springer International Publishing (2016)

[19] Modica, G., Gal, A., Jamil, H.M.: The Use of Machine-Generated Ontologies in Dynamic Information Seeking. Cooperative Information Systems, pp. 433-447, Springer (2001)

[20] Petasis, G., Karkaletsis, V., Paliouras, G., Krithara, A., Zavitsanos, E.: Ontology Population and Enrichment: State of the Art. Knowledge-driven Multimedia Information Extraction and Ontology Evolution, pp. 134-166. Springer-Verlag (2011)

[21] Reyes-Ortiz, J.A., Bravo, M., Pablo, H.: Web Services Ontology Population through Text Classification. Computer Science and Information Systems (FedCSIS), 2016 Federated Conference, IEEE, pp. 491-495 (2016)

[22] Ruiz-Martınez, J.M., Minarro-Giménez, J.A., Castellanos-Nieves, D., Garcıa-Sánchez, F., Valencia-Garcia, R.: Ontology Population: An Application for the E-tourism Domain. Int. J. of Innovative Computing, Information and Control (IJICIC), 7 (11), pp. 6115-6134 (2011)

[23] Uschold, M., Gruninger, M.: Ontologies: Principles, Methods and Applications. The Knowledge Engineering Review, 11 (02), pp. 93-136 (1996)

[24] Vandenbussche, P.Y., Umbrich, J., Matteis, L., Hogan, A., Buil-Aranda, C.: SPARQLES: Monitoring Public SPARQL Endpoints. Semantic Web (Preprint), pp. 1-17 (2016)

[25] Verborgh, R., Vander Sande, M., Hartig, O., Van Herwegen, J., De Vocht, L., De Meester, B., ..., Colpaert, P.: Triple Pattern Fragments: A Low-cost Knowledge Graph Interface for the Web. Web Semantics: Science, Services and Agents on the World Wide Web, 37, pp.184-206 (2016)

[26] Wacker, M.: Linked Data for Cultural Heritage, edited by Ed Jones and Michele Seikel. Chicago: ALA Editions (2016)

Модель семантического поиска на базе тезауруса

© Д.А. Малахов[1] © В.А. Серебряков[1,2]

[1]Московский государственный университет им. М.В. Ломоносова,
[2]Федеральный исследовательский центр «Информатика и управление» РАН,
Москва, Россия

mda.develop@gmail.ru serebr@ultimeta.ru

Аннотация. Представлена модель семантического поиска, которая базируется на применении тезауруса. Описаны ключевые моменты использования модели. Приведены основные возможности тезаурусов, способы их применения в других поисковых системах, а также особенности нашего подхода.

Ключевые слова: семантический поиск, L-теги, применение тезауруса, WordNet.

The Semantic Search Model Based on the Thesaurus

© Dmitriy Malakhov[1] © Vladimir Serebryakov[1,2]

[1]Lomonosov Moscow State University,
[2]Federal Research Center Computer Science and Control of the Russian Academy of Sciences,
Moscow, Russia

mda.develop@gmail.ru serebr@ultimeta.ru

Abstract. This article presents a model of semantic search, based on the thesaurus. The key points of using the model are described. The main features of the thesaurus, the methods of their application in other search systems, and also features of our approach are presented.

Keywords: semantic search, L-tags, thesaurus application, WordNet.

1 Введение

Семантическим поиском, как правило, называется процесс поиска документов по их содержанию. Нетрудно увидеть, что понятие семантического поиска недостаточно формально определено [7]. В частности, понятие содержания или смысла является многозначным.

Существуют различные подходы к реализации семантического поиска. Как правило, выделяют следующие классы моделей семантического поиска:

- Поиск, основанный на структурированных SPARQL запросах к базе знаний в формате OWL/RDF.
- Поиск, основанный на семантическом аннотировании документа с последующей индексацией аннотаций.
- Полнотекстовый поиск, использующий словари синонимов для индексации документов и расширения запросов.
- Всевозможные гибридные решения.

Далее предложена модель семантического поиска, являющаяся гибридным вариантом, так как содержит элементы семантического аннотирования и полнотекстового поиска. В предыдущей работе [2] была представлена модель семантического поиска, основанная на использовании L-тегов.

Определение 1.1. Алфавитом будем называть любое конечное непустое множество. Элементы этого множества называются символами данного алфавита.

Определение 1.2. Термином $t \in T$ алфавита A будем называть любой упорядоченный конечный непустой набор символов алфавита A.

Определение 1.3. L-тегом на множестве терминов T будем называть любой непустой упорядоченный набор терминов из T.

Основная идея использования L-тегов заключается в разбиении алгоритма расчёта релевантности на два этапа. Первый этап заключается в выделении L-тегов в документе и оценке их значимости для этого документа. Эта оценка характеризуется функцией семантики, отображающей пару («документ», «L-тег») в действительное число от 0 до 1. Для качественного поиска важно, чтобы выделенные L-теги полностью покрывали содержание документа. Выделение этого этапа позволяет производить сложные вычисления для расчёта релевантности, причём не во время выполнения запроса, а на этапе индексации.

Второй этап заключается в поиске L-тегов, схожих с запросом пользователя на естественном языке. Поисковый запрос является L-тегом, поэтому релевантность L-тега запросу может быть оценена с помощью функции схожести двух L-тегов,

Труды XIX Международной конференции «Аналитика и управление данными в областях с интенсивным использованием данных» (DAMDID/ RCDL'2017), Москва, Россия, 10–13 октября 2017 года

отображающей пару («запрос», «L-тег») в действительное число от 0 до 1. Функция схожести должна рассчитываться во время выполнения запроса, поэтому должна выполняться достаточно быстро. Комбинация функции схожести и функции семантики характеризует релевантность запроса документу.

Каждый L-тег описывает некоторую информационную потребность. Различные реализации функций семантики и схожести отличаются друг от друга различным пониманием информационной потребности, различным способом оценки схожести информационных потребностей и удовлетворения информационной потребности. Ниже с помощью понятия контекста будут описаны способы реализации функции семантики и функции схожести.

Если в качестве L-тегов рассматривать предложения или абзацы в тексте, то представленная модель позволяет существенно уменьшить поисковый индекс за счет игнорирования L-тегов с достаточно малым значением функции семантики.

Использование модели позволяет применять единые механизмы поиска в случае индексирования не только текстов, но и семантических аннотаций, привязанных к этим текстам, так как семантическую аннотацию можно представить как L-тег или набор L-тегов. Модель семантического поиска в том виде, в котором она была описана ранее, является достаточно общей и не регламентирует, как именно должны быть определены функции семантики и схожести. В настоящей работе предложены уточнения модели семантического поиска для случаев, когда имеется достаточно хороший тезаурус.

2 Применение тезаурусов

2.1 Термины и понятия

К тезаурусам могут быть отнесены достаточно разные словари и лингвистические ресурсы [6]:

- Идеографический словарь, основное назначение которого – помощь в подборе близких по смыслу слов при написании текста.
- Информационно-поисковый тезаурус описывает отношения между терминами предметной области.
- Тезаурус типа WordNet описывает отношения между лексическими значениями естественного языка.
- Ассоциативный тезаурус, описывающий ассоциации людей или совместную встречаемость слов в тексте, рассчитанную автоматически.

Как правило, тезаурусы оперируют двумя сущностями: термином и понятием. Под термином понимается слово или словосочетание, имеющее некоторое смысловое значение. Особенностью естественного языка является то, что одно и то же смысловое значение может быть передано различными терминами. В тезаурусах смысловое значение принято называть понятием, а набор терминов, которые передают это смысловое значение, – синсетом.

Для естественного языка также характерно, что один и тот же термин в разных контекстах характеризует разные понятия. Хороший тезаурус в рамках сферы своего применения должен определять всевозможные понятия термина, а также предоставлять информацию о том, как определить понятие, которое термин характеризует в некотором контексте.

Под тезаурусом мы будем понимать словарь, оперирующий понятиями, которые характеризуются синонимическими рядами (синсетами) и имеют между собой семантические связи, как вертикальные, так и горизонтальные. Далее мы более подробно рассмотрим информационно-поисковый тезаурус и WordNet.

2.2 Информационно-поисковый тезаурус

Информационно-поисковые тезаурусы создавались для описания различных предметных областей и использовались для ручной разметки документов и запросов. Основная идея использования такого рода тезауруса заключалась в определении применяемой терминологии для использования в запросах и индексации документов. Впоследствии эксперименты показали, что эффективность полнотекстового индексирования сравнима с эффективностью поиска, использующего ручное индексирование по [5], [6]. С учетом трудоемкости ручного индексирования оно все чаще заменялось полнотекстовым поиском.

Казалось бы, что информационно-поисковые тезаурусы могут быть полезными в семантическом поиске, но есть две основные проблемы:

- Ориентированность на узкую предметную область не позволяет описать всевозможные значения того или иного термина в целом. В свою очередь документы зачастую могут охватывать различные предметные области, а у пользователей могут быть различные потребности. Поэтому для наиболее полного описания документа может понадобиться несколько тезаурусов, часть понятий которых может пересекаться. В этом случае мы сталкиваемся с проблемой интеграции тезаурусов.
- Более важной проблемой является то, что такие тезаурусы создавались для людей, а не для машин. Поэтому они могут не содержать полного списка синонимов в синсетах, так как подразумевается, что человек догадается, в каком случае нужно привязывать понятие.

Эксперименты по автоматическому индексированию документов и запросов на базе информационно-поисковых тезаурусов не привели к их практическому использованию для автоматической обработки текстов [6].

Таким образом, информационно-поисковые тезаурусы не могут быть использованы явным образом для задачи семантического поиска.

2.3 WordNet

Основные гипотезы, на базе которых разработан WordNet [1]:

- Гипотеза отделимости означает, что лексический уровень языка может быть отделен от морфологического и синтаксического.
- Гипотеза «образца» означает, что существует формальное описание для большинства слов языка.
- Гипотеза о покрытии означает, что словарь должен быть достаточного размера для покрытия всех понятий, чтобы быть эффективным в задачах автоматической обработки текстов.

Разработчики WordNet считают, что два термина могут находиться в одном синсете понятия, если замена одного термина на другой в контексте этого понятия не изменяет смысла предложения. В таком случае термины считаются синонимами. Большинство синсетов имеет толкования. Если термин имеет несколько значений, то он входит в несколько синсетов.

Самые распространенные связи в WordNet:

- Родовидовое отношение используется для существительных. Синсет X называется гипонимом синсета Y, если считается справедливым утверждение: «X – это вид Y». Родовидовое отношение выстраивает иерархию с наследованием всех свойств вышестоящего нижестоящим.
- Отношение «часть–целое» используется для существительных. Синсет X является частью синсета Y, если считается справедливым утверждение: «X – это часть Y».
- Отношение антонимии используется для существительных, прилагательных и наречий, причем связываются не понятия, а термины. Считается, что термины X и Y – антонимы, если одно исключает второе, например, победа – поражение, мужчина – женщина.
- Отношение между однокоренными словами, используется для существительных и различных глагольных форм.

Для описания глаголов были выделены специальные отношения:

- Отношение следования устанавливается между синсетами V1 и V2, если из предложения «Кто-то V1» следует, что «Этот кто-то V2».
- Отношение тропонимии представляет особый вид следования и устанавливает родовидовые отношения между глаголами: «Делать V1 означает делать V2 особым способом».
- Отношение причины связывает два глагольных синсета V1 и V2 следующим образом: «Если кто-то V1, то кто-то другой V2».

Основная критика тезаурусов типа WordNet касается следующих проблем:

- Много значений одного и того же слова. Эту проблему пытались устранить кластеризацией [4], [6].
- Понятия не связаны по контексту. Так называемая «Теннисная проблема». Это усложняет выделение понятия в тексте с разрешением неоднозначности. Эту проблему пытались устранить введением доменов для большинства понятий, где домен характеризует предметную область понятия [3], [6].
- Проблема родовидовых отношений заключается в том, что под этой связью могут скрываться принципиально разные отношения: типы и роли. Эти отношения различаются в способах наследования свойств, поэтому их стоит различать.

Несмотря на критику, WordNet наилучшим образом подходит как тезаурус для задачи семантического поиска, так как наиболее полно описывает понятия и их синсеты. Далее под тезаурусом будем понимать лингвистический ресурс типа WordNet.

3 Контекст

Под контекстом понятия, как правило, подразумевают факторы, влияющие на то, что некоторый термин обозначает некоторое понятие.

Контекст может быть полезен:

- для разрешения неоднозначности термина при выделении понятий в тексте;
- для определения семантической схожести запроса и текста.

Рассмотрим некоторое множество терминов T и множество понятий N.

Определение 3.1. Под абзацем будем понимать группу предложений, идущих в тексте друг за другом, комбинация которых отражает некоторую единую мысль, что приводит к близости контекстов понятий из этого абзаца. Каждое предложение является упорядоченным набором терминов из T, обозначающих некоторые понятия из N.

Исходя из предположения, что контекст понятия характеризуется терминами, которые находятся в одном абзаце с термином понятия, ниже дано формальное определение контекста.

Определение 3.2. Пусть даны множество терминов T и конечное множество абзацев P, где абзац $p \in P$. Вектором абзаца p будем называть вектор действительных чисел V_p размерности $|T|$, компоненты которого соответствуют терминам из T и равны 0 или 1, если термин включен в p или нет, соответственно. Вектор V_p будем считать элементом нормированного векторного пространства, где норма $\|V_p\| = \sqrt{V_p \cdot V_p}$.

Определение 3.3. Пусть даны множество понятий N, множество абзацев P, и для каждого понятия $n \in N$ множество абзацев P_n, в которых было выделено понятие n. Контекстом понятия $n \in N$ будем называть вектор $C_n = (\sum_{p \in P_n} V_p) / |P_n|$.

Таким образом, под контекстом понятия будем понимать среднее арифметическое векторов абзацев, в которых понятие присутствует. Заметим, для того, чтобы посчитать контекст понятия, нужно сначала выделить его в абзацах, причем этих абзацев должно

быть достаточно много, иначе полученный результат будет неустойчивым.

3.1 Выделение контекста понятия

Чтобы определить контекст понятия, нужно выделить это понятие во всех его абзацах, значит, разрешить неоднозначность терминов.

Существуют различные подходы [3], [6] к разрешению неоднозначности при выделении понятия в тексте. Ниже будет предложен альтернативный подход, основной особенностью которого является использование кластеризации.

Сформулируем задачу следующим образом. Даны:

- множество терминов T, множество понятий N;
- множество понятий N_t, обозначенных термином t, для каждого термина $t \in T$;
- для каждого термина $t \in T$ множество абзацев P_t, в которые включен термин t;
- для каждого понятия $n \in N$ множество терминов T_n, описывающих понятие n.

Для каждого понятия $n \in N$ необходимо определить множество абзацев P_n, в которых выделено понятие n.

Для каждого $t \in T$ нужно разбить множество P_t с помощью алгоритма кластеризации, например, k-means++, на $|N_t|$ кластеров. Каждый полученный кластер абзацев P_{t_i} будет характеризовать некоторое понятие N_t. Контекст этого понятия относительно множества абзацев P_{t_i} обозначим как C_{t_i}.

При достаточном объеме данных контексты понятия, построенные по разным группам абзацев, не должны сильно отличаться. Скалярное произведение двух контекстов будет максимальным, если это контексты одного понятия. Поэтому для каждого понятия $n \in N$ нужно собрать все P_{t_i}, такие, что $t \in T_n$. Для каждого термина $t \in T_n$ необходимо оставить только один кластер P_t, так, чтобы оставшиеся кластеры для разных терминов были максимально близки друг к другу. Близость группы кластеров можно оценить с помощью функции $\sum_{t_i \in T_n,\, t_j \in T_n} C_{t_i} \cdot C_{t_j}$.

Для каждого понятия $n \in N$ оставшиеся кластеры P_t объединяются в P_n, по нему считается контекст понятия C_n.

3.2 Выделение понятий в абзаце

При выделении понятий в абзацах мы сталкиваемся с проблемой многозначности и «Теннисной проблемой» (раздел 2.3). Эти проблемы могут быть решены использованием информации о контексте понятия.

Определение 3.4. Пусть даны множество понятий N и множество абзацев P. Степенью близости понятия $n \in N$ и абзаца $p \in P$ будем называть функцию близости

$$affinity(V_p, C_n) = (V_p \cdot C_n)/(\|V_p\| \, \|C_n\|).$$

Рассмотрим абзац $p \in P$ и термин $t \in p$. Пусть N_t – множество понятий, которые могут обозначать термин t. Будем исходить из того, что мы должны выбрать понятие, контекст которого максимально похож на абзац p, тогда в качестве понятия, обозначаемого термином t, следует выбирать n_t, такое, что:

$$affinity(V_p, C_{n_t}) = \max_{n \in N_t} affinity(V_p, C_n).$$

Из-за многозначности может получиться так, что вектор абзаца похож на контексты сразу нескольких понятий. В этом случае предложенный алгоритм может быть улучшен. Мы можем привязать к термину не одно понятие, а несколько, с условием, что контекст каждого привязанного понятия близок к вектору абзаца как минимум на $M\%$ от близости контекста понятия n_t к вектору абзаца p, где M – некоторый порог.

3.3 Выделение понятий в поисковом запросе

Особенностью выделения понятия в поисковом запросе является то, что поисковый запрос в отличие от абзаца имеет намного меньше терминов. Часто поисковый запрос представляет собой последовательность из нескольких терминов, вот почему приведенный выше способ выделения понятий невозможно применить для поисковых запросов.

Определение 3.5. Пусть даны множество терминов T и множество поисковых запросов Q, где запрос $q \in Q$ является набором терминов из T. Вектором запроса q будем называть вектор действительных чисел V_q размерности $|T|$, компоненты которого соответствуют терминам из T и равны 0 или 1, если термин включен в q или нет, соответственно.

Пусть из запроса q каким-то образом было выделено множество понятий N_q. Тогда мы можем дать определение контексту запроса.

Определение 3.6. Пусть даны множество терминов T и множество поисковых запросов Q, где запрос $q \in Q$ является набором терминов из множества T. Контекстом запроса q будем называть вектор

$$C_q = (\sum_{n \in N_q} C_n)/|N_q|.$$

Если пользователь регулярно использует поисковую систему, работая со своими избранными предметными областями, то у нас есть информация о его интересах, и мы могли бы ее использовать.

Исходя из предположения, что контекст пользователя может быть определен через историю его запросов, можно дать следующее определение.

Определение 3.7. Пусть пользователь u последовательно задал K запросов. Контекстом пользователя u будем называть вектор $C_u = \sum_{k=1}^{K} \frac{C_{q_k}}{2^{K-k}}$.

Изначально контекст пользователя представляет собой вектор нулей. После выполнения очередного запроса q контекст уточняется.

На практике контекст пользователя будет разрастаться, то есть будет появляться все больше

ненулевых компонент. Для обнуления наиболее слабых компонент вектора контекста существуют следующие варианты:

- ограничение минимального значения ненулевой компоненты;
- ограничение максимального количества ненулевых компонент.

Определение 3.8. Семантическим ядром запроса q у пользователя u будем называть вектор $S_q = C_u + V_q$.

Выше мы предположили, что множество понятий N_q для запроса q уже выделено, но не описали процесс выделения понятий из запроса. Далее мы исходим из предположения о том, что понятия, выделяемые из запроса, зависят как от запроса, так и от контекста пользователя. Для выделения понятий N_q из запроса q можно воспользоваться алгоритмом выделения понятия абзаца из раздела 3.2. В этом случае вместо вектора абзаца V_p нужно использовать семантическое ядро запроса S_q.

4 Уточнение модели поиска

Использование тезауруса позволяет привязывать понятия как к текстам документов, так и к поисковым запросам. Для этого необходимо посчитать контексты понятий, используя большой массив данных. Далее мы уточним предложенную ранее модель поиска, определив функции схожести поискового запроса, L-тега и семантики L-тега в контексте документа. Будет продемонстрировано, как выделенные понятия и их связи могут быть использованы для семантического поиска.

4.1 Расчёт функции семантики

В качестве L-тегов рассмотрим абзацы документов. Пусть даны конечное множество документов D, где каждому документу d соответствует набор его абзацев P_d, и множество понятий N, для которых предварительно рассчитаны контексты. Задача заключается в вычислении оценки функции семантики $sem(d,p)$ для документа $d \in D$ и абзаца $p \in P_d$, где из абзаца p выделено множество понятий N_p, у каждого понятия $n \in N_p$ есть контекст C_n.

Определение 4.1. Контекстом абзаца p будем называть вектор $C_p = (\sum_{n \in N_p} C_n) / |N_p|$.

Определение 4.2. Контекстом документа d будем называть вектор $C_d = (\sum_{p \in P_d} C_p) / |P_d|$.

Исходя из предположения, что в абзаце выделены все значимые понятия, можно считать, что контекст абзаца C_p характеризует его смысловое значение. Смысловое значение документа определяется смысловым значением его абзацев, что характеризуется контекстом документа C_d. На основании этого может быть определена функция семантики

$$sem(d,p) = \frac{C_d \cdot C_p}{\|C_d\| \, \|C_p\|}.$$

4.2 Расчет функции схожести

Пусть дано множество понятий N. Между этими понятиями существуют родовидовые связи. Функцию близости двух понятий n_1 и n_2 в иерархии будем обозначать $\rho(n_1, n_2)$. В дальнейшем будем считать, что эта функция задана на основе иерархии понятий в тезаурусе, используемом для поиска. Рассмотрим поисковый запрос $q \in Q$ и абзац $p \in P$. Считаем, что в запросе выделены понятия N_q, а в абзаце выделены понятия N_p.

Функция схожести L-тегов должна определять, насколько пересекается смысл, передаваемый L-тегами. Исходя из предположения, что в абзаце и запросе выделены все значимые понятия, а понятия L-тега полностью передают его смысл, можно для запроса q и абзаца p определить функцию схожести

$$sim(q,p) = \sum_{n_1 \in N_q} \frac{\max_{n_2 \in N_p} \rho(n_1, n_2)}{2|N_q|} + \sum_{n_1 \in N_p} \frac{\max_{n_2 \in N_q} \rho(n_1, n_2)}{2|N_p|}$$

4.3 Расчет релевантности

Пусть даны множество запросов Q и множество абзацев P. Рассмотрим запрос $q \in Q$ и абзац $p \in P$. Для расчета релевантности необходимо учитывать:

- $sem(d,p)$ – функция семантики.
- $sim(p,q)$ – функция схожести.

Сначала с помощью функции семантики отбираются похожие на запрос q абзацы P_q. Далее набор P_q сортируется на основе значений функции семантики и функции схожести. Релевантность должна быть больше, если значение функции семантики или схожести больше.

Функция семантики и функция схожести могут быть неравномерно распределены. В этом случае абзацы, которые больше похожи на свои документы, могут получить необоснованное преимущество перед другими абзацами. Чтобы неравномерность функции семантики не приводила к сильному изменению сортировки, можно воспользоваться следующим подходом:

- сортируем P_q по значениям функции схожести, для каждого $p \in P_q$ получаем порядковый номер в отсортированном наборе $simOrder(q,p)$;
- сортируем P_q по значениям функции семантики, для каждого $p \in P_q$ получаем порядковый номер в отсортированном наборе $semOrder(d,p)$;
- релевантность может быть оценена как сумма или произведение $simOrder(q,p)$ и $semOrder(d,p)$.

5 Применение

Рассмотрим поисковый запрос “Java”. О чем пользователь думал, когда задавал этот запрос? Он мог думать о следующем:

- Java – это язык программирования.
- Java – это остров.

- Java – это кофе.

Очевидно, что без использования истории запроса невозможно догадаться о значении термина “Java”, поэтому история запросов является важным компонентом.

Допустим, в истории часто встречается программирование, поэтому к запросу можно привязать понятие «Java – это язык программирования». Пусть в некотором абзаце встречается термин “Java”, если в этом абзаце также встречаются компьютерные термины, то к абзацу на этапе индексирования будет привязано понятие «Java – это язык программирования». В этом случае мы найдем по запросу все абзацы, связанные с языком программирования Java. Полнотекстовый поиск нашел бы все упоминания термина “Java”, но многие абзацы могли бы быть нерелевантными, кроме того, абзацы, в которых нет термина “Java”, но относящиеся к языку программирования Java, не были бы найдены.

Допустим, что по запросу “Java” найдено много абзацев, и все они одинаково похожи на запрос. Как можно ранжировать такую поисковую выдачу? Для этого может быть использована функция семантики. Абзацы, которые лучше передают смысл документа, имеют большую релевантность.

Пусть к некоторым документам вручную привязан L-тег “Java” и определено значение функции семантики. В этом случае L-тег “Java” может участвовать в поиске вместе с другими L-тегами. Привязка поисковых запросов к документам вручную позволяет улучшить качество поиска в наиболее важных темах, кроме того, такой подход используется в рекламных системах.

Представленная модель позволяет вынести сложные вычисления оценки функции семантики на этап индексации, что снижает нагрузку на сервер в момент поиска. Кроме того, появляется возможность контролировать объем поискового индекса и, как следствие, нагрузку на сервер в момент выполнения поискового запроса. Это возможно за счет ограничения количества тегов по значению функции семантики.

6 Заключение

В работе представлена модель семантического поиска и продемонстрирована полезность тезаурусов типа WordNet. Дан небольшой обзор по типам тезаурусов и предложено решение некоторых проблем.

Были формализованы определения контекстов: понятия, абзаца, документа, запроса и пользователя. Были описаны алгоритмы для выделения контекстов с использованием большого корпуса текстов, наиболее полного тезауруса. Была уточнена модель семантического поиска, введенная ранее. Предложены способы оценки функций семантики и схожести с помощью различных контекстов, связей понятий из тезауруса. Была введена, но недостаточно формализована, функция близости понятий. Предполагается ее формализация в дальнейших работах. Кроме того, планируется:

- Описать особенности индексирования математических текстов.
- Рассказать о программной архитектуре, основанной на представленной модели.
- Оценить качество и быстродействие системы поиска по сравнению с другими решениями.

Благодарности

Работа выполнена при финансовой поддержке РФФИ (проект 17-07-00214).

Литература

[1] Fellbaum, C.: WordNet. Blackwell Publishing Ltd, (1998)

[2] Malakhov, D., Sidorenko, Y., Ataeva, O., Serebryakov, V.: Semantic Search in a Personal Digital Library. In: Kalinichenko L., Kuznetsov S., Manolopoulos Y. (eds). Data Analytics and Management in Data Intensive Domains. DAMDID/RCDL 2016. Communications in Computer and Information Science, 706. Springer, Cham (2017)

[3] Magnini, B., Strapparava, C.: Experiments in Word Domain Disambiguation For Parallel Texts. Proc. of the ACL-2000 Workshop on Word Senses and Multi-linguality. Association for Computational Linguistics, pp. 27-33 (2000)

[4] Miller, G.A., Fellbaum, C., Tengi, R.: WordNet. Cambridge, Princeton University (2006)

[5] Salton, G., McGill, M.J.: Introduction to Modern Information Retrieval (1986)

[6] Лукашевич, Н.В.: Тезаурусы в задачах информационного поиска, М.: Изд-во МГУ (2011)

[7] Серебряков, В.А. Что такое семантическая цифровая библиотека In: RCDL 2014. сс. 21-25 (2014)

Development of BWW Ontology for a Workflow Conceptual Modeling

© Igor Fiodorov

Plekhanov Russian University of Economics (PRUE)
Nezhinskaya str., 7, Moscow, Russia

Igor.Fiodorov@mail.ru

Abstract. A success of business process modeling depends on the level of formalization of semantics of a subject domain. The Bunge-Wand-Weber ontology (BWW) is a top-level model containing the most general concepts that are not tied to any specific universe of discourse, on its basis lower-level ontologies relevant to certain domain can be developed. This ontology is often used to evaluate business process modeling languages. Unfortunately the BWW ontology has several shortcomings that limit its practical usage. We adapt the ontology in such a way that it becomes suitable for business processes modeling. We consider we are not allowed to introduce new concepts into BWW ontology, thus we give a new explanation to existing concepts in order to reflect necessary notions.

Keywords: business process modeling, Bunge-Wand-Weber ontology, workflow semantics.

1 Introduction

A variety of languages and notations, namely: UML [1], BPMN [2], EPC [3], ebXML [4], BPEL [5], Petri Nets [6] are used in a workflow so the question emerge to carry out a comparative analysis to determine which is better-suited for business process modeling [7]. These models are often transformed from one language to another, for example, EPC to BPMN and to BPEL, thus second query arises if these notations can equally represent a domain of a discourse? The semantics of these languages is determined in a text form and in some cases ambiguous, thus formalization is required.

Y. Wand and R. Weber hypothesized: if a modeling language or notation is built on top of ontology, then the models created on this basis correctly reflect surrounding world, and is easier to understand [8]. This paper suggests a semiotic approach to prove this supposition. They applied an ontology presented by M. Bunge [9, 10] to a modeling of information systems and suggested a representation model defining a set of constructs that are thought by the Wand and Weber to be necessary and sufficient to describe the structure and behavior of the real world. Thus they have defined a top level ontology that underlies knowledge representation formalism in field of IT development and business process modeling. The ontology is named after author's – Binge-Wand-Weber (BWW) [8]. Unfortunately, this ontology has some shortcomings that limit its practical implementation. Since this is a top-level ontology describing the most general categories, we cannot arbitrarily change the set of its concepts. Therefore, in this paper, we propose to reinterpret some concepts, but staying in a context of original ontology by M. Bunge. Y. Wand and R. Weber supposed that a "good" modeling language can reflect all concepts of the top-level ontology [11]. But we demonstrate that a big number of modeling languages are not capable to map all ontology's concepts thus demonstrate a deficit of expressiveness.

2 Related works

We will regard the business process modeling notation as an artificial language. Ch. Morris distinguished semantics defining a model meaning, syntax determining relations between signs, and pragmatics studying relations between signs and their users [12]. C. Peirce suggested differentiate between textual languages, whose alphabet consists of letters joined into words that convey the meaning, and iconic languages, where each sign denotes a separate notion and provokes emergence of a sensory image [13]. D. Harel and B. Rumpe classed the visual business process modeling languages as iconic, whose alphabet consists of a finite number of graphic signs, each having its own semantic content [14]. They identified relations between the language components, showed that the semantics is defined by means of semantic domain, which lists all the concepts of underlying domain, and semantic mapping connecting the set of modeling language signs with the semantic domain (see Figure 1). Unfortunately, they did not identify the semantic domain scope and the properties of semantic mapping thus restricting the practical use of their approach.

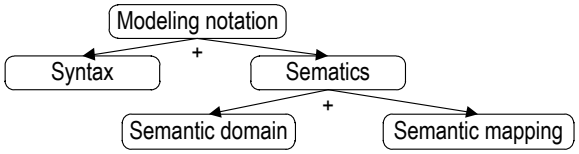


Figure 1 A modeling language semantics [15]

Y. Wand and R. Weber supposed that expressiveness of a modeling language can be assessed as a semantic mapping property. They suggested that the ontological clarity of the modeling language can be evaluated by comparing the alphabet of this language with the

Proceedings of the XIX International Conference "Data Analytics and Management in Data Intensive Domains" (DAMDID/RCDL'2017), Moscow, Russia, October 10–13, 2017

constructs of the top level ontology known as Bunge-Wand-Weber (BWW) [16]. They formalized the semantic domain for business process modeling language and studied semantic mapping.

3 A semiotic approach to evaluate languages of a business process modelling

We base a semiotic approach on the G. Frege's triangle (see Figure 2) illustrating the principle of reality perception by the analyst. It connects the real world object, the respective language sign and the concept abstracting the notion related to the sign. The universe of discourse that we are going to model is formed by objects aka denotata, the aggregate of which is the subject area of modeling. A concept is a certain notion connected with the modeled real world object; it results from the conceptualization procedure. A total of all the concepts form the sematic domain. A sign is a logical name assigned to the respective concept. A set of all signs forms the language alphabet. A model is constructed from the alphabet. Thus a modeling language sign denotes a real world object if there is a concept associated with it, which it its turn abstracts this real world object [17]. Frege's triangle sides can be interpreted as follows: the conceptualization mapping associates each denotatum with a certain concept, the semantic mapping relates the concept to the sign denoting it, the representation mapping relates a sign to a real world's object, it defines a consistency between the model and the original.

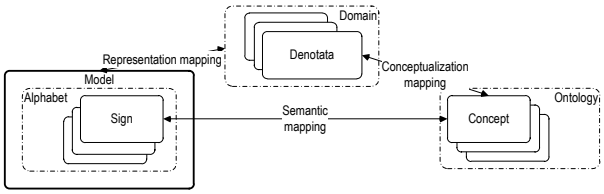


Figure 2 Frege's triangle

Looking at this illustration we can see that a quality of a modeling language can be evaluated by means of a semantic mapping analysis.

4 Semantic mapping

Y. Wand and R. Weber noted that a key success factor of using a given language is its ability to provide the users with a symbol set (modeling primitives), which can directly reflect appropriate ontology concepts (abstracts). They identify the following correspondence options between an alphabet of the modeling language and a set of ontology concepts (Figure 3):

- construct equivalence: each symbol of an alphabet can be associated with exactly one concept;
- construct deficit: separate concepts have no corresponding symbol;
- construct excess: the ontology concept cannot be associated with any symbol;
- construct redundancy (synonymy): one concept can be represented directed in several symbols;
- uncertainty (homonymy): several concepts correspond to one symbol.

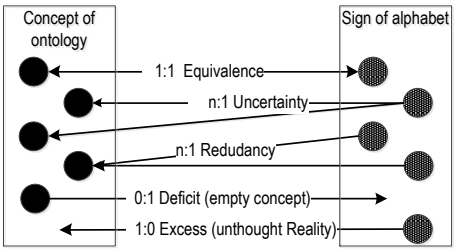


Figure 3 Semantic mapping

The essence of the approach proposed by Y. Wand and R. Weber consists in checking an equivalence of two sets, i.e. symbols of an alphabet and ontology concepts. In the next section we will present the ontology used for business process modeling languages analyses, demonstrate its shortcomings and suggest an extended version.

5 BWW ontological model

The model proposed by Y. Wand and R. Weber is the top level ontology based on ideas developed by M. Bunge [9]. It contains the most general concepts that are not tied to any specific universe of discourse; on its basis lower-level ontologies relevant to certain domain can be developed. Let us discuss this ontology.

M. Bunge calls a thing (aka concrete object) a substantial individual together with its properties. He supposes the world is made up of things. We will treat a thing as a "separate object of the tangible world with relative independence, objectivity and stability of existence" [18], therefore, in what follows the term "object" will be used as a synonym of a thing. Any object has at least one property. A property is an attribute of an object, it can't exist by itself and must be attached to a thing. A property cannot have properties. The object state is defined as a set of all values of all its properties at a given time. Moreover, not all states are considered as acceptable and not all transitions between states are considered lawful [19]. The object state transits due to transformation, which is always implemented by a predetermined rule called the transformation law. Transformation can be interpreted as a work changing the object's properties, or an operation being performed on the object. An event is considered as a change of object's state. In this ontology all and only object change, and every change is an event that is characterized by a pair of an initial and resulting states. A change of a state in the object under investigation is called an internal event, while a change happening in another object which belongs to the environment is considered an external event. Two objects are linked if one of them knows that the others have changed and can react accordingly. Summing up: a process is a history of an object changing its state as a result of a transformations initiated by events. The strength of BWW ontology is in defining a few but really basic concepts, its weakness is in apparent simplicity, leaving a space for misinterpretation.

Here is an example of misconception [19]. State changes can happen either due to internal transformations in thing (self-action of a thing) or due to interactions among different things. Initially the object resides in so called stable state. Due to an action from something outside called an external event the object leaves a stable state and traverses a sequence of unstable states until a new stable state is reached. A process is a sequence of unstable states leading to a stable state. We argue: there are no stable and unstable states as well as self-actions. Being stable means be unchanging. But there is no such property like being changing. We can make a proposition that an object change, but a proposition is not a property. Also, a self- action is impossible because nothing acts upon itself. These misinterpretations had happen because the authors have departed from Bunge's original design.

In our opinion BWW ontology has several shortcomings important for business process modeling:

1. There is no concept for representing an actor participating in process execution. This seems strange and contradicts the established practice to start modeling with the identification of process participants.
2. It remains unclear how to classify a logical operator that route the control flow but do not change a state of an object being processed. While a transformation always changes a state of an object.
3. The ontology does not utilize the category of a time, although it is obvious that the temporal parameters of process execution are very important.

We adapt the ontology in such a way that it can be used to business processes modeling. We consider we are not allowed to introduce new concepts into BWW ontology, thus we give a new explanation to existing concepts in order to reflect necessary notions.

6 Enhanced BWW ontological model

Here and after we restrict our research to an information object only, and keep other types of objects out of discussion. Informational object is a material one because it is stored on a physical medium and is recorded by means of physical principles. Informational object has attributes (aka properties), we associate them with state variables. Now we can say that a process is a trajectory of a phase point in a phase space. We also suppose an object to have a deterministic behavior.

According to M. Bunge a transformation is the only cause of an object change. Wand and Weber consider a transformation by itself, with no connection to other concepts, so it remains unclear what is its origin. We explain a transformation keeping ideas of M. Bunge in mind. He distinguishes between a spontaneous change of an object, for example due to radioactive decay, and induced one, resulting from the interaction between objects. We neglect a spontaneous change because information objects are not subject to aging. Thus a transformation is a result of an interaction between two or more objects. For the simplicity, but not losing a generality of reasoning, we consider one object acting upon another, and the latter does not react back. M. Bunge calls the former an agent and the latter a patient [9]. We will use a term an actor instead of an agent it can be a man or a machine doing a useful work that changes an information object. Thus, a transformation is a useful abstraction separating a unit of work from an actor performing it.

Y. Wand and R. Weber suppose that any transformation result in a change of an object's state. That is why remains unclear how classy logical operators (gateway) because they do not change a state of information object. Let us pay attention to the fact that M. Bunge differentiates between the intrinsic object properties inherent thereto and distinguishing one entity instance from another one (for example, the color and shape characterize each object on an individual basis) and mutual properties, which characterize one object relative to another (for example, distance is a property of a pair of objects). Speaking about the transformation, Wand and Weber have in mind a change of intrinsic properties only. We will interpret the transformation in a more comprehensive sense, and also consider a change of mutual properties. For example, an operation changes the intrinsic properties of the object, while the logical operator in a process diagram route the object along one of several processing paths, changing its relative position, whereas the intrinsic properties of the object remain unchanged. Therefore, by partitioning the transformations which change the intrinsic properties of the object and the transformations which modify the mutual properties, we complement the ontology with a capability to represent logical process operators [20].

Y. Wand and R. Weber define an event as a fact of changing an object state, irrespective of the cause of occurrence. Meanwhile, it remains unclear what is the difference between an event and a state. In the existing interpretation, the term "event" means a change of a state, it makes sense "for the reason of this" and reflects the cause-effect relationship – the next operation can start because of the completion of the previous one.

The interpretation of an event proposed by us is different from the interpretation suggested by M.Bunge. By the definition of E. Babkin, an event is something that is happening at some instant per saltum, step-wise and is considered as a state change of a certain object [21]. Yu. Pavlovsky interprets an event as an instant in time designating a change of the object states [22]. Therefore, we will link an event with a moment in time when a change of state of a certain object occurred; it has the meaning of "afterwards" – later in the chronological order. Thus, an internal event establishes the fact and the moment in time when the object passed into the following state and is ready for execution of the next operation. The occurrence of an internal event is insufficient for the beginning of execution of the next operation. In case of an interactive operation the execution begins following the interference of an actor and the latter is treated as an external object relative to the system. If the operation is automatic, then it starts after a signal from an external control device. Therefore, external event represents the fact and the moment in time

of changing the state of the object outside of the system, which initiates the execution of the operation and record the moment when the transformation began. Thereby, the terms of temporal logic are added to the ontology: a moment in time and time interval between two consecutive events [20]. The time interval between the occurrence of an internal event indicating readiness to processing, and an external event indicating the real beginning of work will be interpreted as the waiting time, the time interval between the occurrence of an external event indicating the beginning of work and internal event indicating the end of processing will be interpreted as the execution time. An external event not only initiates the execution of the process operation, but can also coordinate it. For example, a customer placed an order – this event initiates the process, and if the customer canceled the order, further processing may not be reasonable. The external event may imply the occurrence of an abnormal situation and require special processing.

Figure 4 shows an object that changes its state after a current transformation. This state transition is considered an internal event, now the object is ready for a following transformation and is waiting for an external event. An external event reflects a change of a state in another object, belonging to the environment, it initiates transformation and a cycle repeat. We claim that for transformation the internal event is necessary but not sufficient, a next transformation starts only after an external event. A logical operator (a gateway) can start without an external event.

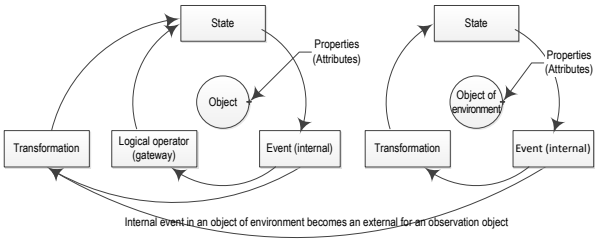


Figure 4 Linked objects

The new concept of an event is treated in accordance with the representations of temporal logic. A time interval between an external event initiating a transformation and external event canceling a transformation is associated with duration of a transformation, while another time interval between an external event canceling a previous transformation and external event initiating a next transformation and is associated with a waiting time before it can start.

This approach allows us to explain, analyze and identify errors in business process models when an external event occurs earlier than an internal event associated with it. For example, in most cases the next possible error goes unnoticed: the external event occurs earlier than the object will go into the corresponding state and starts waiting for an event. Since the external event is not remembered, it will be lost. Therefore, an object achieving a desired state, will wait indefinitely for an event that has already happened in the past.

Thus, we enhanced the BWW ontology, added an actor – a man or a machine who performs a useful work; separated transformations: those changing intrinsic properties correspond to operations, others changing mutual properties correspond to logical operators; changed the interpretation of an event concept such that it designates a fact and a moment in time when the object state changes; we also demonstrated that the external events are related to each process operation. An important conclusion that can be made from the analysis of BWW enhanced ontology is in specifying a set of concepts:

- the object to be processed – it has an internal structure describing a set of inherent properties of the object;
- transformations changing intrinsic properties of the object that result in a change of its state, that are mapped to a process operations;
- transformations changing relative properties and thus route the object, are mapped to a gateways;
- internal events designate a moment in time when the object is ready for execution of the next operation;
- external events designate a moment in time when external actor starts operation.
- a transformation can start only in case both preconditions fulfilled: a previous transformation is successfully finished and external event happen.
- time intervals between an external and internal events characterize a duration of transformation and waiting time before it can start.
- any transformation must be mapped to an actor performing a useful work.

7 Evaluation of semantics by means of the BWW ontology

As an example, we consider the EPC notation. Methods ARIS 7.0 defines four main elements of notation: functions, events, connections and rules. A function is called a “subject-oriented task or an action performed on an object” [23]. Let's associate a function and a transformation. An event in the EPC is called “the fact that the information object has received the status associated with the business process”. Events “switch functions” i.e. transfer control from one function to another [23]. That means we should associate an event and a state of the object being processed. This explanation is valid for intermediate and terminate events but is not legal for a start event. A start event represent a terminate state of the object that belongs another process that was executed prior to evaluated process. A connection “defines the logical links between the objects they connect”. We can consider a connection both, as a trajectory of functions or as a history of object’s states. There is no explicit definition of a tern “rule” but we can understand that a rule routes a process execution. Notation introduces two kinds of rules: function and event ones. A function rule selects next operation to be performed while an even rule can be considered as a condition on a state transition diagram that governs a permissible change of a state. Thus, we can see that EPC model is a combination of a state transition and data flow diagrams.

In a second example we shortly consider a BPMN notation [24]. An official specification defines an activity as a work that is performed within a business process so we can associate it with a transformation. An activity can be atomic or compound a latter is called a process. Here we find inconsistency: according to BWW a process is a trajectory of an object in its phase space while BPMN interprets a process as a sequence or flow of activities in an organization with the objective of carrying out work. A similar contradiction can be found in definition of a sequence flow. To facilitate a discussion, BPMN specifications employ the concept of a token that traverse the sequence flows and pass through activities. A token is introduced as a theoretical concept that is used as an aid to define the behaviour of a process that is being performed by describing how activities interact with a token. The only suitable BWW concept to represent a token would be an information object itself, but this consideration will require a total rethink of BNPM semantics. Last but not least, an event in BPMN is similar to an event in BWW, first it represent a transformation that listen for a notification coming from other object and after it arrive perform a useful work. Second it represents a transformation that monitor an object and after it change its state it sends a notification to other objects

This short example demonstrates only principles of using BWW ontology to evaluate semantics of a particular modeling notation and find contradictions or inconsistencies. BWW also helps to establish a mapping between several notations to simplify a port of a model. We found that semantics of signs having equal name but belonging to different languages can be dissimilar which means we can't directly map, for example, EPC event to BPMN events. It can be seen that both languages are capable to represent different number of BWW concepts thus they have different expressive power. Let us investigate the expressiveness in detail.

8 The analysis of business processes modeling languages

A large body of research reveals that process modeling languages and notations are not capable of reflecting BWW ontological model concepts all at once, but only part of them. Moreover, the authors of investigations focus their attention on a percentage ratio of modeled and unmodeled concepts, calculate a relative degree of deficit, redundancy, excess and overload. Table 1 shows the results of similar research [25]. One is compelled to ask: to what extent a language having a 10% of deficit is better than another language having a 15% expressiveness deficit?

Let us suggest that a requirement of equivalence of language symbols set and BWW ontology concepts is too strict, that the overload, redundancy and excess make the modeling language unsuitable for modeling. However, the expressiveness deficit of the language is acceptable, because it can be overcome. Table 1 shows a comparison of the EPC and BPMN expressive power in order to represent various perspectives of the process model. Both notations do not model the structure of information object; thus, they do not reflect the information perspective. The symbol "event" in EPC notation reflects a state acquired by an object as a result of execution of the process operation. It makes it possible to show a sequence of state transitions and thus model objects behavior; however, no place for state mapping is foreseen in BPMN notation. Both notations represent names of the operations which transform the information object, but it is necessary to refine them using mini-specifications, to specify the properties to be changed in order to achieve a target state. The EPC diagram contains no means to indicate time intervals; therefore, it does not represent a temporal perspective – such means are available in BPMN notation. Both diagrams enable us to reflect logical process statements. In summary, it can be seen that none of the business process modeling notations are able to represent the process model perspectives all at once, but only part of them. In other words, both notations have an expressiveness deficit.

Table 1 Analysis of notations expressiveness

Notation	Relative level of			
	Deficit	Redun-dancy	Excess	Over-load
BPMN 1.0	51%	35%	28%	25%
BPML 1.0	29%	65%	28%	3%
EPC	3%	62%	43%	28%
WSCI 1.0	29%	49%	18%	8%
ebXML 1.01	15%	13%	14%	5%
BPEL 1.1	32%	49%	13%	6%

9 Conclusions

The research demonstrates that BWW ontology is a strong tool for defining a modeling languages axiomatics. First it serves to express modeling language semantics. By mapping a sign of a language on concept of ontology we are able to specify a precise and unified description that helps to avoid any kind of misunderstandings about its meaning. Second by evaluating the mapping of set signs onto set of ontology concept one can make a judgment on the overall quality of a modeling language.

The novelty of this research is manifested first in an adaptation of BWW ontology for process modeling languages analyses. Due to a specific of a task we are not allowed to introduce any new concepts into top-level ontology, that is why we reinterpret some concepts, staying strictly in a context of original ontology by M. Bunge. The major achievement is in introduction of notions of time, actor, logical operator. Second we propose a semiotic approach to evaluate business process modeling languages, thus proving a hypothetical assertion by Wand and Weber that a language can be evaluated by mapping on BWW ontology. In this paper we apply this method to investigate language semantics, but it opens a new opportunity for evaluation of language grammar and pragmatics as well.

A practically important result is obtained, proving that none of the known business process modeling languages is capable to represent all BWW ontological

concepts at once, but only part of them. Thus, all known modeling notations have an expressiveness deficit. This gives grounds to conclude that the process modeling should be carried out simultaneously in several notations, so that each particular model showed a limited set of properties of the simulated phenomenon, and all together they gave a complete and comprehensive picture of the simulated reality.

The ideas and methods presented in this research can be also applied to improve axiomatics of the object oriented programming. For example OOP terms: object, attribute, method, message are declared without precise explanation. Easy to see their direct correspondence to BWW concepts, which opens a way for rigorous specification of the semantics.

References

[1] Opdahl A. Henderson-Sellers B. Ontological evaluation of the UML using BWW model // Software and Systems Modeling, Vol. 1, No. 1, 2002. pp. 43-67

[2] Recker J. Rosemann M. Krogstie J. Ontology-Versus Pattern-Based Evaluation of Process Modeling Languages: a Comparison // Communications of the Association for Information Systems, Vol. 48, No. 20. pp. 774-799

[3] Green P. Rosemann M. Integrated Process Modeling. An Ontological Evaluation // Information Systems, Vol. 25, No. 2, 2000. pp. 73-87

[4] Green P. Rosemann M. Indulska M. Ontological Evaluation of Enterprise Systems Interoperability huUsing ebXML // IEEE Transactions on Knowledge and Data Engineering, Vol. 17, No. 5, 2005. pp. 713-725

[5] Green P. Rosemann M. Indulska M. Manning C. Candidate Interoperability Standards: An Ontological Overlap Analysis // Data & Knowledge Engineering, Vol. 62, No. 2, 2007. pp. 274-291.

[6] Rosemann M. Green P. Indulska M. Recker J. Using ontology for the representational analysis of process modelling techniques // International Journal of Business Process Integration and Management, Vol. 4, No. 4, 2009. pp. 251-265

[7] Fiodorov I. Comparative analysis of business processes modeling methods // Open Systems. 2011. No. 8. pp. 28-30

[8] Wand Y., Weber R. An Ontological Model of an Information System // IEEE Transactions on software engineering, Vol. 16, No. 11, 1999. pp. 1282-1292

[9] Bunge M. Treatise on Basic Philosophy Ontology I: The Furniture of the World. Vol 3. Boston, MA: D. Reidel Publishing Company, 1977. 369 pp.

[10] Bunge M. Treatise on Basic Philosophy Ontology II:The World of Systems. Vol 3. Boston, MA,: D. Reidel Publishing Company, 1979

.

[11] Wand Y. Weber R. Toward a theory of the deep structure Of information systems // Journal of Information Systems, Vol. 5, No. 3, 1995. pp. 203-223

[12] Morris C. Signification and significance: a study of the relations of signs and values. M.I.T. Press, Massachusetts Institute of Technology, 1964. 99 pp.

[13] Peirce C.S. Peirce on Signs: Writings on Semiotic. Chapel Hill, NC: University of North Carolina Press, 1991. 294 pp.

[14] Harel D., Rumpe B. Modeling Languages: Syntax, Semantics and All That Stuff, Part I: The Basic Stuff, Weizmann Science Press of Israel©, Jerusalem, Israel, Technical Report 2000. 1-28 pp.

[15] Harel D., Rumpe B. Meaningful Modeling What's the Semantics of Semantics // Journal Computer, Vol. 37, No. 10, October 2004. pp. 64-72

[16] Wand Y., Weber R. Research Commentary: Information Systems and Conceptual Modeling -- A Research Agenda // Information Systems Research, Vol. 13, No. 4, 2002. pp. 363-376

[17] Ullmann S. Semantics: An Introduction to the Science of Meaning. Oxford,: Basil Blackwell, 1972. 278 pp.

[18] Uemov A.I. Veshchi, svoystva i otnosheniya [Things, properties and relations]. (in Russian). ed. Moscow: USSR Academy of Sciences, 1963. 183 pp.

[19] Soffer P. Wand Y. On the Notion of Soft-Goals in Business Process Modeling // Business Process Management Journal, Vol. 11, №. 6, 2005. pp. 663 - 679.

[20] Fiodorov I.G. Adaptatsiya ontologii Bunge-Vanda-Vebera k opisaniyu ispolnyaemykh modeley biznes-protsessov [Adaptation of Bunge-Wand-Weber ontology to description of executable business processes models] // Applied Informatics, Vol. 58, No. 4, 2015. pp. 82–92

[21] Babkin E.A. O ponyatii sobytiya v diskretno-sobytiynom modelirovanii [On the concept of event in discrete-event modeling] // In: Information Systems. Theory and Practice. Kursk: Kursk State University, 2010. pp. 46–51

[22] Pavlovskiy Y.N., Belotelov N.V., Brodskiy Y.I. Imitatsionnoe modelirovanie [Simulation modeling]. (In Russian) Moscow: 2008. 237 pp.

[23] Software AG. ARIS Method v.7, Darmstadt, 2011.

[24] Business Process Model and Notation (BPMN) v 2.0. OMG. (2011) https://www.elma-bpm.ru/bpmn2/

[25] Recker J., Rosemann M., Indulska M., Green P. Business Process Modeling: A Maturing Discipline? // BPMcenter.org. 2005. BPM Center Report

Интеграция неоднородных баз данных

Heterogeneous database integration

Спецификация и реализация разномодельных правил интеграции данных

© С.А. Ступников

Институт проблем информатики ФИЦ «Информатика и управление» РАН,
Москва, Россия

sstupnikov@ipiran.ru

Аннотация. Рассмотрен подход к спецификации правил интеграции данных с использованием рекомендации W3C – логического диалекта RIF-BLD. Это позволяет использовать в одном правиле сущности из разных коллекций, представленных в разных моделях данных. Логическая семантика RIF-BLD также позволяет однозначным образом интерпретировать спецификации рассматриваемых правил интеграции. Предложен подход к реализации правил RIF-BLD в языке HIL: это позволяет компилировать правила интеграции в программы вычислительной модели MapReduce и исполнять их в распределенных инфраструктурах, основанных на Hadoop.

Ключевые слова: интеграция данных, модели данных, логические правила, реализация правил.

Specification and Implementation of Multimodel Data Integration Rules

© Sergey Stupnikov

Institute of Informatics Problems, Federal Research Center “Computer Science and Control” of the Russian Academy of Sciences,
Moscow, Russia

sstupnikov@ipiran.ru

Abstract. The paper considers an approach for specification of data integration rules using RIF-BLD logic dialect that is a recommendation of W3C. This allows us to reference entities defined in different collections represented using different data models in the same rule. Logical semantics of RIF-BLD provides also unambiguous interpretation of data integration rules. The paper proposes an approach for implementation of RIF-BLD rules using HIL language. Thus data integration rules are compiled into MapReduce programs and can be executed over Hadoop-based distributed infrastructures.

Keywords: data integration, data models, logic rules, rule implementation.

1 Введение

Хранилища данных в настоящее время являются одной из основных составляющих инфраструктур поддержки систем бизнес-аналитики. Данные извлекаются из различных коллекций, преобразуются к схеме хранилища, интегрируются и загружаются в хранилище. Над хранилищем выстраивается слой приложений, осуществляющих анализ данных (например, при помощи методов математической статистики, машинного обучения и др.) и выдающих результат пользователю в необходимой форме.

Ввиду роста неоднородности источников данных, их количества и объемов данных, актуальными остаются вопросы разработки методов спецификации и реализации интеграции данных в масштабируемых инфраструктурах распределенного хранения и обработки данных, подобных Apache Hadoop [1].

Традиционные промышленные решения по интеграции данных и разработке хранилищ (например, IBM InfoSphere Information Server [2]) предлагают визуальные средства проектирования потоков работ интеграции данных, оперирующие преимущественно понятиями и операциями реляционной модели и порождающие программы трансформации и интеграции данных на языке SQL. Отдельно выделяются визуальные средства сопоставления схем исходных коллекций данных и схемы хранилища (целевой схемы), например, InfoSphere FastTrack [2], позволяющие автоматически порождать части потоков работ интеграции данных. Для обеспечения интеграции данных, представленных в различных моделях данных, такие средства предоставляют отдельные компоненты, позволяющие преобразовывать

Труды XIX Международной конференции «Аналитика и управление данными в областях с интенсивным использованием данных» (DAMDID/ RCDL’2017), Москва, Россия, 10–13 октября 2017 года

исходные данные к реляционной модели (например, компонент преобразования XML в IBM InfoSphere DataStage [3]).

Параллельно промышленным решениям проводятся исследования методов формальной спецификации правил интеграции данных. Классическим примером работы в данном направлении является подход *обмена данными* [9]. Этот подход позволяет описывать интеграцию данных реляционной модели с использованием логических правил, которым сообщается формальная семантика в логике предикатов первого порядка.

В настоящей статье рассматривается подход к спецификации правил интеграции данных с использованием рекомендации W3C – логического диалекта RIF-BLD [6]. RIF-BLD является диалектом каркаса логических диалектов RIF-FLD формата обмена правилами RIF [5], нацеленного на унификацию синтаксиса и семантики языков логических правил. RIF-BLD включает достаточно широкий спектр возможностей спецификации, в частности, позиционные термы и термы с именованными аргументами (обобщающие понятие терма в логике первого порядка), фреймовые термы (выражающие утверждения о структуре объектов), классификационные термы, термы равенства, внешние термы (использующиеся для ссылок на сущности, рассматриваемые как «черные ящики» в пределах спецификации). Это позволяет использовать в одном правиле сущности из разных коллекций, представленных в разных моделях данных. Рассматриваются правила, в голове которых могут присутствовать лишь предикаты, соответствующие сущностям целевой схемы, а в теле – предикаты, соответствующие сущностям исходных схем.

Логическая семантика RIF-BLD позволяет однозначным образом интерпретировать спецификации рассматриваемых правил интеграции и допускает их реализацию с использованием различных языков – от декларативных (например, SQL) до императивных (например, Java). В данной работе рассматривается подход к реализации логических правил RIF-BLD в языке высокого уровня HIL [10], [7], разработанного компанией IBM и поставляемого в составе Hadoop-решения BigInsights 3.0 [11], а также как часть InfoSphere Big Match for Hadoop [12]. Распределенное исполнение программ на HIL в среде Hadoop достигается путем их компиляции в программы на Java, которые, в свою очередь, исполняются с использованием средств поддержки в Hadoop вычислительной модели MapReduce [14].

Подходы к спецификации и реализации правил интеграции иллюстрируются на примерах интеграции неоднородных коллекций данных по Арктической зоне в хранилище информации, нацеленной на поддержку поисково-спасательных операций. В рамках проекта «Извлечение информации из разноструктурированных данных для решения задач информационной поддержки поисковых действий в арктической зоне» (РФФИ, 15-29-06045) были, в частности, выбраны неоднородные коллекции данных, подлежащие интеграции [16], и разработана единая схема хранилища [20]. В данной статье в качестве примеров рассматриваются коллекции, представленные в реляционной модели, XML, документной модели MongoDB, графовой модели Neo4j. Схема хранилища представлена в реляционной модели.

Структура статьи выглядит следующим образом. В разделах 2–4 проиллюстрированы подход к спецификации правил интеграции данных, представленных в различных моделях (раздел 2 – XML и реляционная модель, раздел 3 – документная модель, раздел 4 – графовая модель) с использованием диалекта RIF-BLD и подход к реализации правил в языке HIL. В разделе 5 обобщены применяемые принципы спецификации и реализации правил интеграции.

2 Спецификация и реализация правил интеграции данных XML и реляционной модели с разрешением конфликтов

В данном разделе представлены два примера правил интеграции данных, оперирующих сущностями XML и реляционной модели.

В первом примере рассмотрено правило преобразования данных из XML к реляционной схеме хранилища (целевой схеме) с разрешением структурных конфликтов, конфликтов имен и значений.

Во втором примере рассмотрено правило преобразования данных из соединения двух коллекций, одна из которых представлена в XML, другая – в реляционной модели.

2.1 Интеграция данных о маршрутах объектов

В левом столбце Таблицы 1 приведен пример данных в формате XML о маршруте объекта, полученных из системы мониторинга, учета и классификации судов КИИС «МоРе» [18]. Маршрут (элемент *ISSKOI_Track*) состоит из маршрутных точек (*ISSKOI_TrackPoint*). В каждой точке заданы значения координат (*pos*), времени (*Time*) и высоты (*BarAltitude*). Данные о других маршрутах имеют ту же структуру элементов и отличатся значениями элементов и атрибутов.

Таблица 1 Данные о маршруте объекта и соответствующие элементы целевой схемы

Пример данных в исходной модели (XML)	Элементы целевой схемы (реляционная модель)
<ISSKOI_Track> <Id>56473</Id> <TrackName>copter-1</TrackName> <ISSKOI_TrackPoints> <ISSKOI_TrackPoint id="uuid-2b7ca14"> <Position> <Point id="uuid-859bef91">	Track(PK trackId, name) TrackPoint(PK pointId,

<pos>33.8957 246.37</pos> </gml:Point> </Position> <Time>2016-12-12 T13:33:11</Time> <BarAltitude>533.89 </BarAltitude> <HSpeed>108.1</HSpeed> <VSpeed>2</VSpeed> </TrackPoint> </TrackPoints> </Track>	FK path, time, height, latitude, longitude)

В правом столбце таблицы приведены элементы целевой схемы, соответствующие исходным данным. Так, элемент *ISSKOI_Track* соответствует отношению целевой схемы *Track*, вложенный элемент *Id* соответствует первичному ключу (PK) *Track.trackId,* вложенный элемент *TrackName* – атрибуту *Track.name.* Элемент *ISSKOI_TrackPoint* соответствует отношению целевой схемы *TrackPoint*, атрибут элемента *id* соответствует первичному ключу *TrackPoint.pointId,* вложенные элементы *Time* и *BarAltitude* – атрибутам *TrackPoint.time* и *TrackPoint.height* соответственно, вложенный элемент *pos* – атрибутам *TrackPoint.logtitude* и *TrackPoint.latitude*.

В рассмотренном примере, как и в других примерах, приведенных ниже, приведена лишь часть элементов, составляющих исходные данные и целевую схему.

Очевидно, что при спецификации правила интеграции данных о маршрутах необходимо разрешить конфликты имен (например, элемент *BarAltitude* и атрибут *height* имеют одинаковую семантику, но разные имена), структурные конфликты и конфликты значений (элемент *pos*, содержащий широту и долготу, соответствует двум отдельным атрибутам *logtitude, latitude*).

С использованием диалекта RIF-BLD необходимое правило интеграции может быть описано следующим образом:

```
   Forall ?Track ?Id ?TrackName ?TrackPoints
?TrackPoint ?Position ?Time ?Height ?Point ?pos
   ( AND( Track(trackId->?Id  name->?TrackName)
     TrackPoint(pointId->?pid  path->?Id
      time->?Time  height->?Height
      latitude->External(get_latitude(?pos))
      longitude->External(get_longitude(?pos))
))
   :-
   AND(?Track#ISSKOI_Track
       ?Track[Id->?Id  TrackName->?TrackName
             TrackPoints->?TrackPoint]
     ?TPoint#ISSKOI_TrackPoint
     ?TPoint[id->?pid  Position->?Position
       Time->?Time  BarAltitude->?Height]
     ?Position#Position        ?Position[Point-
>?Point]
     ?Point#Point  ?Point[pos->?pos]))
```

Forall в правиле обозначает квантор всеобщности, знак «:-» обозначает импликацию – логическое следование формулы-головы правила из формулы-тела правила, операция *AND* обозначает конъюнкцию. Идентификаторы вида *?X* обозначают переменные.

В правиле использованы три вида предикатов. В голове правила применены *предикаты с именованными аргументами* [6] *Track* и *TrackPoint*, соответствующие отношениям целевой схемы. В теле правила использованы *предикаты членства* [6] (например, предикат *?Track#ISSKOI_Track*, обращающийся в истину, когда переменная *?Track* принимает значение произвольного элемента *ISSKOI_Track*) и *фреймовые предикаты* [6], отражающие структуру XML-элементов исходных данных. Так, предикат *?Point[pos->?pos]* обращается в истину на таких парах значений переменных *?Point* и *?pos*, что элемент – значение переменной *?Point* имеет вложенный элемент *pos* и его значение есть *?pos*. Конъюнкция предикатов в теле правила полностью задает структуру вложенных элементов и атрибутов произвольного элемента *ISSKOI_Track*.

Связь значений атрибутов отношений в голове правила и значений элементов и атрибутов в теле правила задается при помощи переменных, таким образом разрешаются конфликты имен и структурные конфликты.

Конфликты значений могут быть разрешены при помощи функций (например, *get_latitude*), представляемых в RIF-FLD как *внешние термы* (*External*). Их семантика в рамках RIF-FLD напрямую не определяется и уточняется при реализации (например, семантика функции *get_latitude* состоит в том, чтобы вернуть первое число, входящее в исходную строку).

Рассмотренное правило имеет естественную логическую семантику: для всех наборов значений подкванторных переменных, обращающих в истину все предикаты тела правила на исходной коллекции, отношения хранилища должны содержать кортежи, соответствующие предикатам головы правила с тем же набором значений переменных. Для рассмотренного примера хранилище должно содержать кортежи *Track(trackId: 56473, name: “copter-1”), TrackPoint(pointId: “uuid-2b7ca14”, path: 56473, time: “2016-12-12T13:33:11”, height: 533.89, latitude: 33.8957, longitude: 246.37).*

Вышеописанное правило может быть реализовано в языке HIL следующей программой:

```
   declare ISSKOI_Track: ?;
   declare Track: ?;
   declare TrackPoint: ?;
   declare get_latitude:
     function string to double;
   declare get_longitude:
     function string to double;

   @jaql{
   get_latitude = fn($s) convert(substring(
     $s, 0, strPos($s, ' ')-1), schema double);
   get_longitude = fn($s) convert(substring(
     $s, strPos($s, ' ')+1, strLen($s)),
     schema double);
   }
   insert into Track
   select [ trackId: t."Id",name: t."TrackName"]
   from ISSKOI_Track t;

   insert into TrackPoint
   select [ path: t."Id", time: p."Time",
     height: tpt."BarAltitude",
     hSpeed:        tpt."HSpeed",        vSpeed:
tpt."VSpeed",
     course: tpt."Course",
```

```
    latitute: get_latitude(pt."pos"),
    longitude: get_longitude(pt."pos") ]
  from ISSKOI_Track t, t."TrackPoints" tpt,
    tpt."Position" ps, ps."Point" pt;
```

В программе объявляются (*declare*) исходная сущность (внешний элемент *ISSKOI_Track*) и целевые сущности, соответствующие отношениям (*Track, TrackPoint*). Знак «?» в объявлении означает, что структура сущностей не задана при определении, а выводится из программы.

Объявляются функции разрешения конфликтов (*get_latitude, get_longitude*) и их сигнатуры. Реализация функций производится с использованием языка Jaql [4] (директива *@jaql*) и его функций работы со строками *substring, strPos.*

Для целевых отношений *Track, TrackPoint* определяются операторы *insert*, порождающие кортежи этих отношений. В секции *from* операторов *insert* объявляются исходные сущности, каждому предикату членства в теле правила RIF-BLD (например, *?Track#ISSKOI_Track*) соответствует объявление с точностью до имени переменной (например, *ISSKOI_Track t*).

Атрибуты целевых отношений и выражения порождения их значений определяются в секции *select* в соответствии с предикатами в голове правила и фреймовыми предикатами в теле правила. Например, предикату головы *Track(trackId->?Id)* и фреймовому предикату тела *?Track[Id->?Id]* соответствует определение *trackId: t."Id"* в секции *select* оператора *insert into Track.*

Заметим, что язык HIL оперирует данными в формате JSON [13], поэтому для применения рассмотренного правила на языке HIL для преобразования данных о маршрутах в целевую схему необходимо преобразовать исходные XML-документы в JSON при помощи встроенной функции *xmlToJson* языка Jaql [4]. Полученные на выходе HIL-программы JSON-документы затем загружаются в реляционное хранилище над Hadoop (например, *Hive* [15]).

2.2 Интеграция данных о судах

В левом столбце Таблицы 2 приведены пример данных о судах в формате XML, полученный из системы «Поиск-Море» (*ERRTableShips*) [19]; а также пример данных о судах в реляционной модели, полученный из системы ЕСИМО [17] (данные получены в формате CSV и преобразованы в JSON). В обеих исходных коллекциях могут быть найдены данные об одних и тех же судах (судно идентифицируется по названию и позывному). Кроме того, коллекции содержат различные взаимодополняющие данные о судах.

В правом столбце таблицы приведены элементы целевой схемы, соответствующие исходным данным. Данные, соответствующие судну, как транспортному средству, сосредоточены в отношении *Vessel*; как спасательной единице – в отношении *SARUnit*; как юридической сущности – в отношении *LegalEntity*.

Таблица 2 Данные о судах и соответствующие элементы целевой схемы

Пример данных в исходных моделях (XML, реляционная)	Элементы целевой схемы (реляционная модель)
<ERRTableShips> <Id>64694571</Id> <Name>мвс Ростов Великий</Name> <Callsing>UBZG5</Callsing> <Dates> <StartDate>2017-01-08</StartDate> <EndDate>2017-01-09</EndDate> <Dates> </ERRTableShips> [{ "Platforma_nazvanie": "мвс Ростов Велкий", "Strana_naimenovanie": "Россия", "Organizaciya_nazvanie": "ФГУП Балтийское БАСУ – Сахалинский филиал", "Pozyvnoj": "UBZG5", "nomer_IMO": "9586796" }]	SARUnit(beginDuty, endDuty, vehicle) Vessel(PK vehicleId, name, call, country, FK owner, imoNumber) LegalEntity(PK entityId, name)

С использованием диалекта RIF-BLD необходимое правило интеграции описано следующим образом:

```
Forall ?Id ?name ?call ?beginDuty ?endDuty
  ?country ?ownerName ?imoNumber
( Exists ?owner (
  AND( SARUnit(beginDuty->?beginDuty
       endDuty->?endDuty  vehicle->?Id )
    Vessel(vehicleId->?Id
      name->External(normalize(?nazv))
      call->?call  Country->?country
      owner->?owner   imoNumber->?imoNumber)
)
    LegalEntity(entityId->?owner
      name->?ownerName) ))
:-
AND( ?ERRTableShips#ERRTableShips
  ?ERRTableShips[Id->?Id  Name->?name
    Callsing->?call  Dates->?Dates]
  ?Dates#Dates
  ?Dates[StartDate->?beginDuty
         EndDate->?endDuty]
  Ship("Platforma_nazvanie"->?nazv
   "Pozyvnoj"->?call
   "Strana_naimenovanie"->?country
   "Organizaciya_nazvanie"->?ownerName
   "nomer_IMO"->?imoNumber)
  External(compareShipName(?name,   ?nazv))
))
```

Аналогично примеру, приведенному в предыдущем подразделе, в голове правила конъюнкцией соединяются предикаты, соответствующие отношениям целевой схемы. Кроме того, конъюнкция в голове правила заключена под квантор существования (*Exists*) по переменной *?owner*[18], значение которой (не определенное в исходных данных) является первичным ключом

[18] Строго говоря, квантор всеобщности в голове правила выходит за границы диалекта RIF-BLD, однако входит в каркас RIF-FLD.

entityId в кортеже отношения *LegalEntity* и внешним ключом в кортеже отношения *Vessel*.

В теле правила конъюнкцией соединяется предикат *Ship*, соответствующий отношению из системы ЕСИМО, предикаты членства и фреймовые предикаты, соответствующие структуре XML-документов из системы «Поиск-Море».

Отдельной особенностью правила является то, что при соединении сущностей из исходных коллекций происходит проверка на соответствие имен судов с некоторой точностью (возможны варианты записей и ошибки) при помощи функции *compareShipName*.

Вышеописанное правило может быть реализовано в языке HIL следующей программой (объявления сущностей и функций опущены):

```
create link ShipLink as
select [
 Callsing_Name:
 [Callsing: es.Callsing, Name: es.Name],
 "Pozyvnoj_Platforma_nazvanie":
 ["Pozyvnoj": s."Pozyvnoj",
 "Platforma_nazvanie":
s."Platforma_nazvanie"]]
from ERRTableShips es, Ship s
match using
 rule1: es.Callsing = s."Pozyvnoj" and
   compareShipName(es.Name,
     s."Platforma_nazvanie");

insert into SARUnit
select [ vehicle: s.Id,
         beginDuty: d.StartDate,
         endDuty: d.EndDate
       ]
from ShipLink sl, ERRTableShips s, s.Dates d
where sl.Callsing_Name.Name = s.Name and
  sl.Callsing_Name.Callsing = s.Callsing;

insert into Vessel
select [ vehicleId: s.Id,
 name: normalize(s."Platforma_nazvanie"),
 call: s."Pozyvnoj",
 country: s."Strana_naimenovanie",
 owner: get_id(s."Organizaciya_nazvanie"),
 imoNumber: s."nomer_IMO"]
from ShipLink sl, Ship s
where      sl.Callsing_Name.Callsing      =
s."Pozyvnoj"
   and  sl.Callsing_Name.Name =
          s."Platforma_nazvanie";

insert into LegalEntity
select [
 entityId:
get_id(s."Organizaciya_nazvanie"),
 name: s."Organizaciya_nazvanie"]
from ShipLink sl, Ship s
where sl.Callsing_Name.Name =
      s."Platforma_nazvanie"  and
  sl.Callsing_Name.Callsing = s."Pozyvnoj";
```

Соединение сущностей исходных коллекций производится с использованием оператора разрешения сущностей *create link*. В секции *from* оператора указываются соединяемые коллекции (*ERRTableShips*, *Ship*), в секции *select* – составные ключи (*Callsing_Name*, *Pozyvnoj_Platforma_nazvanie*), однозначно идентифицирующие исходные сущности, правило сопоставления сущностей *rule1* (совпадение позывных и соответствие имен с точностью до функции *compareShipName*).

Как и в примере из предыдущего подраздела, для каждого отношения целевой схемы в программе определен оператор *insert*. Особенность данной программы состоит в том, что целевые отношения пополняются на основании только тех сущностей, которые связаны оператором *create link*.

Квантор всеобщности реализуется с использованием функции *get_id*, порождающей уникальный идентификатор и тем самым означающей подкванторную переменную *?owner*. Значение порожденного идентификатора присваивается первичному ключу *LegalEntity.entityId* и внешнему ключу *Vessel.owner*.

3 Спецификация и реализация правил интеграции данных документной модели

В данном разделе рассмотрен пример правила интеграции данных, оперирующего сущностями документной модели. Исходная коллекция содержит сообщения о происшествиях в Арктической зоне из социальных сетей и сущности (персоны, суда, географические локации и т. д.), извлеченные средствами анализа текстов из сообщений [8]. Данные хранятся в документной СУБД MongoDB и экспортируются для дальнейшей интеграции в файлах в формате JSON.

В левом столбце таблицы 3 приведен пример данных о сообщении (*Messages*) и данных о сущностях, извлеченных из сообщения (*Entites*). Связь сообщений и сущностей, извлеченных из них, установлена на основании значения составного ключа *id*. В правом столбце таблицы приведены элементы целевой схемы, соответствующие исходным данным.

Таблица 3 Данные о сообщениях и соответствующие элементы целевой схемы

Пример данных в исходной модели (документная)	Элементы целевой схемы (реляционная модель)

{"Messages": [{
"id": {"coll_id":
"8002",
"res_id": {
"site_id":
"9b290c9f3bda",
"doc_id":
"3649a5559a62"}},
"annotation":
"Chinese seismic vessel aimed for Russian #Barents Sea oil at logistics port #Kirkenes",
"metafields": {
"mf203": "2016-4-30",
"mf205": "eng",
"mf200":
"iceblogger"}
}]}
 	
{"Entities": [{
"id": {"coll_id":
"8002",
	Document(
	documentId,
	collection,
	source,
	content,
	time,
	language,
	author)
	
	ExtractedEntity(
	entityId,
	document,
	token,
	tag,
	begin
	end)

"res_id": { "site_id": "9b290c9f3bda", "doc_id": "3649a5559a62"}}, "entities": [{ "s_token": "Barents Sea", "s_tag": "I-ALOC", "s_end": 53, "s_begin": 42}, { "s_token": "Kirkenes", "s_tag": "I-ALOC", "s_end": 85, "s_begin": 77}]}] }	

С использованием диалекта RIF-BLD правило интеграции данных о сообщениях может быть описано следующим образом:

```
Forall ?m ?ext ?doc ?coll ?src ?cont
  ?time ?lang ?auth ?mid ?mres ?mf ?eid ?eres
( Exists ?ent( AND(
 Document(documentId->?doc collection->?coll
  source->?src content->?cont time->?time
  language->?lang author->?auth)
 ExtractedEntity(entityId->?ent document->?doc
  token->?tok tag->?tag begin->?beg
  end->?end) ))
:-
AND( ?m#Messages ?ext#Entities
 ?m[id->?mid annotation->?cont metafields->?mf]
 ?mid[coll_id->?coll res_id->?mres]
 ?mres[site_id->?src doc_id->?doc]
 ?mf[mf203->?time mf205->?lang mf200->?auth]
 ?ext[id->?eid entities->?ents]
 ?eid[coll_id->?coll res_id->?eres]
 ?eres[site_id->?src doc_id->?doc]
 ?ext#?ents
 ?ext[s_token->?tok s_tag->?tag
      s_begin->?beg s_end->?end] ))
```

Аналогично примерам, приведенным в предыдущем разделе, в голове правила конъюнкцией соединяются предикаты, соответствующие отношениям целевой схемы. В теле правила при помощи предикатов членства и фреймовых предикатов отражена структура документов, соответствующих сообщениям и извлеченным сущностям. Правило может быть реализовано в языке HIL следующей программой:

```
insert into Document
select [ documentId: mres."doc_id",
  collection: mid."coll_id",
  source: mres."site_id",
  content: m."annotation",
  time: mf."mf203",
  language: mf."mf205",
  author: mf."mf200"]
from Message m, m."id" mid, mid."res_id" mres,
     m."metafields" mf;

insert into ExtractedEntity
select [ entityId: get_id(strcat(
            eres."site_id",
eres."doc_id")),
  document: eres."doc_id",
  token: e."s_token",
  tag: e."s_tag",
  begin: e."s_begin",
  end: e."s_end"]
from Entities ext, ext."id" eid,
     eid."res_id" eres, ext.entities e;
```

Для обоих отношений целевой схемы в программе определен оператор *insert*.

4 Спецификация и реализация правил интеграции данных графовой модели

В данном разделе рассмотрен пример правила интеграции данных, оперирующего сущностями графовой модели. Исходная коллекция содержит данные о спасательных операциях, данные хранятся в графовой СУБД Neo4j и экспортируются для дальнейшей интеграции в файлах в формате JSON.

В левом столбце Таблицы 4 приведен пример данных о спасательных операциях. Данные содержат три вида вершин (*Nodes*): спасательная операция (*SarOperation*), координационный центр (*CoordCenter*), координатор операции (*Coordinator*); и два вида ребер (*Relationships*): *rCoordCenter* (ребра, связывающие операцию и центр), *rCoordinator* (ребра, связывающие операцию и ее координатора).

В правом столбце таблицы приведены элементы целевой схемы, соответствующие исходным данным.

Таблица 4 Данные о спасательных операциях и соответствующие элементы целевой схемы

Пример данных в исходной модели (графовая)	Элементы целевой схемы (реляционная модель)
{"Nodes": [{"id":"12", "labels":["SarOperation"], "properties":{ "OperationID":"5824", "OperationDate": "2016-09-15T00:00:00", "OperationName": "Arctic Sunrise Sinking", "countrycode":"643"} }, {"id":"13", "labels":["CoordCenter"], "properties":{ "CoordCenterID":"mrcc667340", "Name":"MRCC Dikson"} }, {"id":"14", "labels":["Coordinator"], "properties":{ "Name": "Schurov V. A.", "CoordinatorID": "5773"} }]} { "Relationships": [{"id":"4", "type":"rCoordCenter", "startNode":"12", "endNode":"13"}, {"id":"9", "type":"rCoordinator", "startNode":"12", "endNode":"14"}] }	SAROperation (PK operationId, FK coordCenter, FK coordinator, country, creationTime, name) CoordinationCenter(PK centerId, name) Person(PK personId, name)

С использованием диалекта RIF-BLD интеграция данных о спасательных операциях может быть описана совокупностью следующих правил:

```
Forall ?pid ?name ?lbl ?prp(
```

```
Person(personId->?pid  name->?name) :-
AND( ?n#Nodes
  ?n[labels->?lbl  properties->?prp]
  "Coordinator"#?lbl
  ?prp[CoordinatorID->pid  Name->?name]))

Forall ?cid ?name ?lbl ?prp(
CoordinationCenter(centerId->?cid
  name->?name) :-
AND( ?n#Nodes
  ?n[labels->?lbl  properties->?prp]
  "CoordCenter"#?lbl
  ?prp[CoordCenterID->cid  Name->?name]))

Forall (
SAROperation(operationId->?opid
  creationTime->?time  name->?name
  country->?country  coordCenter->?cntr
  coordinator->?coord)
:-
AND( ?n#Nodes  ?n[id->?id]
  ?n[properties->?prp  labels->?lbl]
  ?r1#Relatioships  ?r2#Relatioships
  ?n1#Nodes  ?n1[id->?id1]
  ?n2#Nodes  ?n2[id->?id2]
  "SarOperation"#lbl
  ?r1[type->"rCoordCenter"
     startNode->?id  endNode->?id1]
  ?r2[type->"rCoordinator"
     startNode->?id  endNode->?id2]
))
```

Каждому отношению целевой схемы соответствует свой тип вершин, например, отношению *SAROperation* соответствуют вершины с меткой *"SarOperation"* (атрибут *labels*). Поэтому для каждого отношения удобно определить свое порождающее правило. Предикаты, соответствующие отношениям целевой схемы, составляют головы правил. В телах правил при помощи предикатов членства и фреймовых предикатов отражены структуры данных, соответствующие вершинам и ребрам графа.

Правила могут быть реализованы в языке HIL следующей программой:

```
insert into Person
select [
  personId: prp."CoordinatorID",
  name: prp."Name"]
from Nodes n, n."properties" prp,
     n."labels" lbl
where
  array_includes(lbl, "Coordinator");

insert into CoordinationCenter
select [
  centerId: prp."CoordCenterID",
  name: prp."Name"]
from Nodes n, n."properties" prp
     n."labels" lbl
where
  array_includes(lbl, "CoordCenter");

insert into SAROperation
select [
  operationId: prp."OperationID",
  creationTime: prp."OperationDate",
  name: prp."OperationName",
  country: prp."countrycode",
  coordCenter: prp1."CoordCenterID",
  coordinator: prp2."CoordinatorID"]
from Nodes n, n."properties" prp,
  n."labels" lbl,
  Relationships r1, Relationships r2,
  Nodes n1, Nodes n2
where
  array_includes(lbl, "SarOperation") and
  n."id" = r1."startNode" and
  r1."type" = "rCoordCenter" and
  n1."id" = r1."endNode" and
  n."id" = r2."startNode" and
  r2."type" = "rCoordinator" and
  n2."id" = r2."endNode";
```

Каждое из правил представляется отдельным оператором *insert*. Предикаты проверки типа вершины (например, *"Coordinator"#?lbl*) реализуются с использованием функции *array_includes:*

```
@jaql{ array_includes = fn(a, e)
 if (not exists(a->filter $ == e)) false
 else true; }
```

5 Общие принципы спецификации правил интеграции данных с использованием RIF-BLD и их реализации в языке HIL

Основные принципы спецификации правил интеграции данных с использованием RIF-BLD, примененные в данной работе, можно обобщить следующим образом:

- правила интеграции имеют вид импликации, посылка которой называется *телом*, а заключение – *головой*;
- голова и тело правил представляют собой формулы, связывающие предикаты операциями конъюнкции или дизъюнкции;
- в теле правила допускаются предикаты, соответствующие сущностям только исходных схем, в голове – только целевой схеме;
- для описания свойств кортежей, принадлежащих отношениям реляционной модели, в правилах используются предикаты с именованными аргументами [5];
- для описания свойств структур данных произвольной степени вложенности (при помощи которых представляются данные различных нереляционных моделей – XML, документной, графовой и т. д.) используются фреймовые предикаты и предикаты членства [5];
- связь значений атрибутов сущностей исходных и целевой схем осуществляется при помощи переменных;
- свободные переменные, использующиеся в теле правила, связываются внешним квантором всеобщности;
- свободные переменные в голове правила, не встречающиеся в теле (фактически, они соответствуют данным, не определенным явно в исходных коллекциях), связываются квантором существования;
- нетривиальные предикаты-условия (отличные от равенства) и функции разрешения конфликтов определяются как внешние термы [5].

Основные принципы реализации правил интеграции данных на RIF-BLD в языке HIL, примененные в данной работе, можно обобщить следующим образом:

- сущности исходных и целевой схем, функции разрешения конфликтов и нетривиальные предикаты-условия объявляются при помощи директивы *declare*; для функций определяется их сигнатура;
- функции реализуются на языке Jaql в отдельных секциях программы, либо на языке Java во внешних файлах;
- для каждой сущности (отношения) в голове правила создается отдельный оператор *insert*, порождающий кортежи этих отношений:
  - в секции *from* объявляются исходные сущности, каждому предикату с именованными переменными или предикату членства в теле правила соответствует объявление с точностью до имени переменной;
  - в секции *select* указываются атрибуты целевых отношений (в соответствии с их названиями в предикатах головы правила) и выражения порождения их значений; выражения формируются в соответствии с термами в предикатах головы правила либо с термами во фреймовых предикатах тела правила;
  - в секции *where* указываются предикаты, отражающие способы соединения исходных сущностей (формируемые на основании совпадения имен переменных во фреймовых предикатах тела) и их отбора (соответствующие предикатам, налагающим условия на данные отдельных исходных коллекций в теле правила);
- если соединение сущностей исходных коллекций в теле правила осуществляется с использованием нетривиальных предикатов и функций (отличных от равенства атрибутов), предварительное соединение сущностей производится с использованием оператора разрешения сущностей *create link*:
  - в секции *from* указываются соединяемые коллекции;
  - в секции *select* указываются составные ключи, однозначно идентифицирующие исходные сущности;
  - в секции *match* указываются правила сопоставления сущностей;
  - коллекция пар сопоставленных сущностей затем используется в качестве входной в операторах *insert*, порождающих кортежи целевых отношений.

6 Заключение

Представлен подход к спецификации правил интеграции данных с использованием рекомендации W3C – логического диалекта RIF-BLD. Широкий спектр возможностей спецификации RIF-BLD позволяет использовать в одном правиле сущности из разных коллекций, представленных в разных моделях данных. Рассмотрены правила, в голове которых могут присутствовать лишь предикаты, соответствующие сущностям целевой схемы, а в теле – предикаты, соответствующие сущностям исходных схем. Логическая семантика RIF-BLD позволяет однозначным образом интерпретировать спецификации рассматриваемых правил интеграции и допускает их реализацию с использованием различных языков. Представлен подход к реализации логических правил RIF-BLD в языке высокого уровня HIL, разработанного компанией IBM. Путем компиляции программ на HIL в программы вычислительной модели MapReduce достигнуто распределенное исполнение правил интеграции в среде Hadoop.

К дальнейшим направлениям работы можно отнести вопросы интерпретации и реализации логических программ на RIF-BLD, в телах правил которых допускаются предикаты целевой схемы. Это требует адаптации, в частности, процедуры погони [9]. Другим направлением является реализация автоматического преобразования спецификаций RIF-BLD в программы, компилируемые и исполняемые на распределенных вычислительных инфраструктурах (например, языке HIL).

Поддержка

Работа выполнена при поддержке РФФИ (гранты 15-29-06045, 16-07-01028).

Литература

[1] Apache Hadoop Project. 2016. http://hadoop. apache.org/

[2] Ballard, C., Alon, T., Dronavalli, N., Jennings, S., Lee, M., Toratani, S.: IBM InfoSphere Information Server Deployment Architectures.ibm.com/ redbooks (2012)

[3] Bar-Or, A., Choudhary, S.: Transform XML using the DataStage XML stage. IBM developerWorks (2011)

[4] Beyer, K.S., Ercegovac, V., Gemulla, R., Balmin, A., Eltabakh, M., Kanne, C.-C., Ozcan, F., Shekita, E.J.: Jaql: A Scripting Language for Large Scale Semistructured Data Analysis. 37th Int. Conf. on Very Large Data Bases VLDB, pp. 1272-1283. Curran Associates, New York (2011)

[5] Boley, H., Kifer, M. (eds.): RIF Framework for Logic Dialects. W3C Recommendation, 2nd edn. (February 5, 2013)

[6] Boley, H., Kifer, M. (eds.): RIF Basic Logic Dialect. W3C Recommendation, 2nd edn. (February 5, 2013)

[7] Burdick, D., Hernández, M.A., Ho, H., Koutrika, G., Krishnamurthy, R., Popa, L., Stanoi, I.R., Vaithyanathan, S., Das, S.: Extracting, Linking and Integrating Data from Public Sources: A Financial Case Study. IEEE Data Eng. Bull., 34 (3), pp. 60-67 (2011)

[8] Devyatkin, D., Shelmanov, A.: Text Processing Framework for Emergency Event Detection in the Arctic Zone. Kalinichenko L., Kuznetsov S., Manolopoulos Y. (eds). Data Analytics and Management in Data Intensive Domains.

DAMDID/RCDL 2016. Communications in Computer and Information Science, 706, pp. 74-88. Springer (2017)

[9] Fagin, R., Kolaitis, P., Miller, R., Popa, L.: Data exchange: semantics and query answering. Theoretical Computer Science, 336 (1), pp. 89-124 (2005)

[10] Hernandez, M., Koutrika, G., Krishnamurthy, R., Popa, L., Wisnesky, R.: HIL: A High-level Scripting Language for Entity Integration. 16th Conf. (International) on Extending Database Technology Proceedings EDBT 2013, pp. 549-560 (2013)

[11] IBM InfoSphere BigInsights Version 3.0 Information Center. https://goo.gl/lZpEQd

[12] InfoSphere Big Match for Hadoop. Technical Overview. https://goo.gl/0TMqvw

[13] Introducing JSON. http://www.json.org/

[14] Miner, D.: MapReduce Design Patterns: Building Effective Algorithms and Analytics for Hadoop and Other Systems. O'Reilly Media (2012)

[15] The Apache Hive data warehouse software. http://hive.apache.org/

[16] Брюхов, Д.О., Скворцов, Н.А., Ступников, С.А.: Методы интеграции разноструктурированных данных по Арктической зоне для извлечения информации, нацеленной на поддержку поисково-спасательных операций. Системы высокой доступности. 13 (2). М.: Радиотехника (2017) (в печати)

[17] Единая государственная система информации об обстановке в мировом океане. http://portal.esimo.ru/portal

[18] Комплексная интегрированная информационная система «МоРе». http://www.marsat.ru/ciis-more

[19] Программный комплекс «Поиск-Море». http://map.geopallada.ru/

[20] Скворцов, Н.А., Брюхов, Д.О.: Разработка схемы хранилища информации для поддержки поисковых действий в Арктической зоне. Системы высокой доступности, 13 (2). М.: Радиотехника (2017) (в печати)

On an Approach to Data Integration: Concept, Formal Foundations and Data Model

© Manuk G. Manukyan

Yerevan State University,
Yerevan, Armenia

mgm@ysu.am

Abstract. In the frame of an extensible canonical data model a formalization of data integration concept is proposed. We provide virtual and materialized integration of data as well as the possibility to support data cubes with hierarchical dimensions. The considered approach of formalization of data integration concept is based on the so-called content dictionaries. Namely, by means of these dictionaries we are formally defining basic concepts of database theory, metadata about these concepts, and the data integration concept. A computationally complete language is used to extract data from several sources, to create the materialized view, and to effectively organize queries on the multidimensional data.

In memory of Garush Manukyan, my father.

This work was supported by the RA MES State Committee of Science, in the frames of the research project N 15T-18350.

Keywords: data integration, mediator, data warehouse, data cube, canonical data model, OPENMath, grid file, XML.

1 Introduction

The emergence of a new paradigm in science and various applications of information technology (IT) are related to issues of big data handling [21]. The concept of big data is relatively new and involves the growing role of data in all areas of human activity beginning with research and ending with innovative developments in business. Such data is difficult to process and analyze using conventional database technologies. In this connection, the creation of new IT is expected in which data becomes dominant for new approaches to conceptualization, organization, and implementation of systems to solve problems that were previously considered extremely hard or, in some cases, impossible to solve. Unprecedented scale of development in the big data area and the U.S. and European programs related to big data underscore the importance of this trend in IT.

In the above discussed context the problems of data integration are very actual. Within of our approach to data integration an extensible canonical model has been developed [16]. We have published a number of papers that are devoted to the investigation of data virtual and materialized data integration problems, for instance [15, 17]. Our approach to data integration is based on the works of the SYNTHESIS group (IPI RAS) [2, 9-12, 22-25], who are pioneers in the area of justifiable data models mapping for heterogeneous databases integration. To support materialized integration of data during creation of a data warehouse a new dynamic index structure for multidimensional data was proposed [6] which is based on the grid files [18] concept. We consider the concept of grid files as one of the adequate formalisms for effective management of big data. Efficient algorithms for storage and access of that directory are proposed in order to minimize memory usage and lookup operations complexities. Estimations of complexities for these algorithms are presented. In fact, the concept of grid files allows to effectively organize queries on multidimensional data [5] and can be used for efficient data cubes storage in data warehouses [13,19]. A prototype to support the considered dynamic indexation scheme has been created and its performance was compared with one of the most demanded NoSQL databases [17].

In this paper a formalization of the data integration concept is proposed using a mechanism of the content dictionaries (similarly ontologies) of the OPENMath [4]. Subjects of the formalization are the basic concepts of database theory, metadata about these concepts and the data integration concept. The result of the formalization are a set of content dictionaries, constructed as XML DTDs on the base of OPENMath and are used to model the databases concepts. With this approach, schema of an integrated database is an instance of content dictionary of the data integration concept. Within the considered approach is provided virtual and materialized integration of data as well as the possibility to support data cubes with hierarchical dimensions. Using OPENMath as the kernel of the canonical data model allows us to use a rich apparatus of computational mathematics for data analysis and management.

The paper is organized as follows: Concept and formal foundations of the considered approach to data integration are presented briefly in Section 2. Canonical data model and issues to support the data integration concept are considered in Section 3. The conclusion is provided in Section 4.

2 Brief Discussion on Data Integration Approach

The basis of our concept to data integration is based on the idea of integrating arbitrary data models. Based on this assumption our concept of data integration assumes:

- applying extensible canonical model;
- constructing justifiable data models mapping for heterogeneous databases integration;
- using content dictionaries.

Choosing the extensible canonical model as integration model allows integrating arbitrary data sources. As we allow integration of arbitrary data sources a necessity to check mapping correctness between data models arises. It is reached by formalization of data model concepts by means of AMN machines [1] and using B-technology to prove correctness of these mappings.

The content dictionaries are central to our concept of data integration and semantical information of different types can be defined based on these dictionaries. The concept of content dictionaries allows us to extend the canonical model by means of introducing new concepts in these dictionaries easily. In other words, canonical model extension *only* is reduced to adding new concepts and metadata about these concepts in content dictionaries. Our concept to data integration is oriented as virtual and materialized integration of data as well as to support data cubes with hierarchical dimensions. It is important that in all cases we use the same data model. The considered data model is an advanced XML data model which is a more flexible data model than relational or object-oriented data models. Among XML data models, a distinctive feature of our model is that we use a computationally complete language for data definition. An important feature of our concept is the support of data warehouses on the base of a new dynamic indexing scheme for multidimensional data. A new index structure developed by us allows to organize effectively OLAP-queries on multidimensional data and can be used for efficient data cubes storage in data warehouses. Finally, the modern trends of the development of database systems lead to use of different divisions of mathematics to data analysis. Within of our concept to data integration, this leads to the use of corresponding content dictionaries of the OPENMath.

2.1 Formal Foundations

The above discussed concept to data integration is based on the following formalisms:

- canonical data model;
- OPENMath objects;
- multidimensional indexes;
- domain element calculus.

Below we will consider these formalisms in detail. As we noted, our approach to data integration is based on the works of the SYNTHESIS group. According to the research of this group, each data model is defined by syntax and semantics of two languages, data definition language (DDL) and data manipulation language (DML). They suggested the following principles of synthesis of the canonical model:

- **Principle of axiomatic extension of data models**

The canonical data model must be extensible. The kernel of the canonical model is fixed. Kernel extension is defined axiomatically. The extension of the canonical data model is formed during the consideration of each new data model by adding new axioms to its DDL to define logical data dependencies of the source model in terms of the target model if necessary. The results of the extension should be equivalent to the source data model.

- **Principle of commutative mappings of data models**

The main principle of mapping of an arbitrary resource data model into the target one (the canonical model) could be reached under the condition that the diagram of DDL (schemas) mapping and the diagram of DML (operators) mapping are commutative.

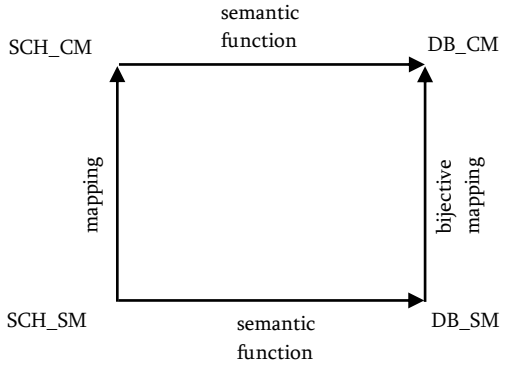


Figure 1 DDL mapping diagram

In Figure 1 we used the following notations: **SCH_CM**: Set of schemas of the canonical data model; **SCH_SM**: Set of schemas of the source data model; **DB_CM**: Database of the canonical data model; **DB_SM**: Database of the source model.

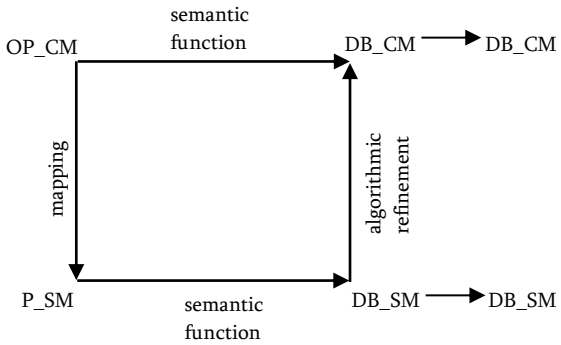


Figure 2 DML mapping diagram

In Figure 2 we used the following notations: **OP_CM**: Set of operators of the canonical data model; **P_SM**: Set of procedures in DML of the source model.

- **Principle of synthesis of unified canonical data model**

The canonical data model is synthesized as a union of extensions.

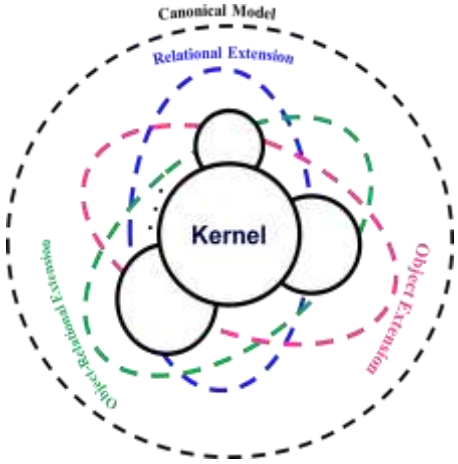


Figure 3 Canonical data model

2.2 Mathematical Objects Representation

The OpenMath is a standard for representation of the mathematical objects, allowing them to be exchanged between computer programs, stored in databases, or published on the Web. The considered formalism is oriented to represent semantic information and is not intended to be used directly for presentation. Any mathematical concept or fact is an example of mathematical object. The OpenMath objects are such representation of mathematical objects which assume an XML interpretation.

Formally, an OpenMath object is a labeled tree whose leaves are the *basic* OpenMath objects. The compound objects are defined in terms of *binding* and *application* of λ-calculus [8]. The type system is built on the basis of types that are defined by themselves and certain recursive rules, whereby the compound types are built from simpler types. To build compound types the following type constructors are used:

- *Attribution*. If *v* is a basic object variable and *t* is a typed object, then a*ttribution* (*v*, *type t*) is a typed object. It denotes a variable with type *t*.
- *Abstraction*. If *v* is a basic object variable and *t*, *A* are typed objects, then *binding* (λ, *attribution* (*v*, *type t*), A) is a typed object.
- *Application*. If *F* and *A* are typed objects, then *application* (*F*, *A*) is a typed object.

The OPENMath is implemented as an XML application. Its syntax is defined by syntactical rules of XML, its grammar is partially defined by its own DTD. Only syntactical validity of the OPENMath objects representation can be provided on the DTD level. To check semantics, in addition to general rules inherited by XML applications, the considered application defines new syntactical rules. This is achieved by means of introduction of content dictionaries. Content dictionaries are used to assign formal and informal semantics to all symbols used in the OPENMath objects. A content dictionary is a collection of related symbols encoded in XML format. In other words, each content dictionary defines symbols representing a concept from the specific subject domain.

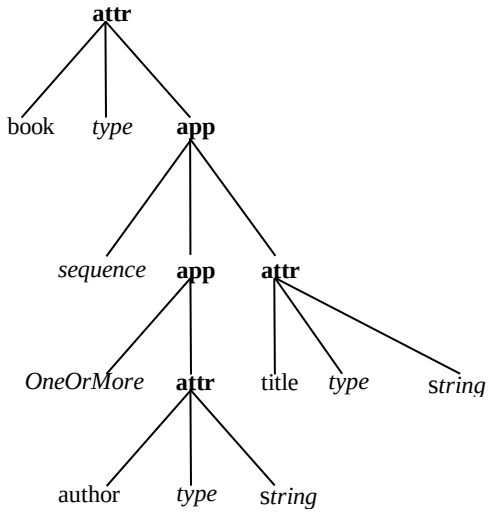


Figure 4 An example of compound object

2.3 Dynamic Indexing Scheme for Multidimensional Data

To support the materialized integration of data during the creation of a data warehouse and to apply very complex OLAP-queries on it a new dynamic index structure for multidimensional data was developed (see more details in [6]). The considered index structure is based on the grid file concept. The grid file can be represented as if the space of points is partitioned into an imaginary grid. The *grid lines* parallel to axis of each dimension divide the space into *stripes*. The number of grid lines in different dimensions may vary, and there may be different spacings between adjacent grid lines, even between lines in the same dimension. Intersections of these stripes form cells which hold references to data buckets containing records belonging to corresponding space partitions.

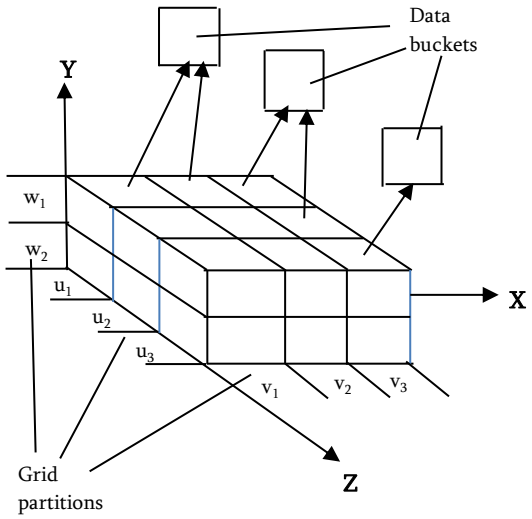


Figure 5 An example of 3-dimensional grid file

The weaknesses of the grid file formalism concept are non-efficient memory usage by groups of cells referring to the same data buckets and the possibility of having a large number of overflow blocks for each data buckets.

In our approach, we made an attempt to eliminate these defects of the grid file. Firstly, we introduced the concept of the chunk: set of cells whose corresponding records are stored in the same data bucket (represented by single memory cells with one pointer to the corresponding data buckets). Chunking technique is used to solve the problem of empty cells in the grid file.

Secondly, we consider each stripe as a linear hash table which allows increasing the number of buckets more slowly (for each stripe, the average number of overflow blocks of chunks crossed by that stripe is less than one). By using this technique we essentially restrict the number of disk operations.

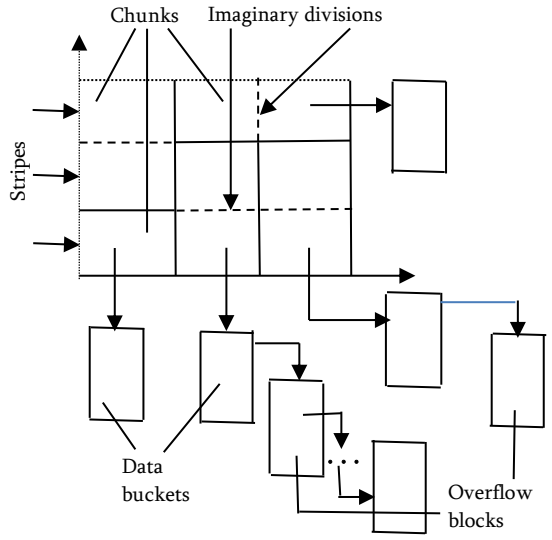


Figure 6 An example of 2-dimensional modified grid file

We perform comparison of directory size by our approach with two techniques for grid file organization proposed in [20]: MDH (multidimensional dynamic hashing) and MEH (multidimensional extendible hashing). Directory sizes for both of these techniques are: $\boldsymbol{O}\left(\boldsymbol{r}^{\mathbf{1}+\frac{\mathbf{1}}{\boldsymbol{s}}}\right)$ and $\boldsymbol{O}\left(\boldsymbol{r}^{\mathbf{1}+\frac{\boldsymbol{n}-\mathbf{1}}{\boldsymbol{ns}-\mathbf{1}}}\right)$ correspondingly, where r is the total number of records, s is the block size and n is the number of dimensions. In our case the directory size can be estimated as $O\left(\frac{nr}{s}\right)$. Compared to MDH and MEH techniques, the directory size in our approach is $\frac{\boldsymbol{sr}^{\frac{\mathbf{1}}{\boldsymbol{s}}}}{\boldsymbol{n}}$ and $\frac{\boldsymbol{sr}^{\frac{\boldsymbol{n}-\mathbf{1}}{\boldsymbol{ns}-\mathbf{1}}}}{\boldsymbol{n}}$ times smaller correspondingly. We have implemented a data warehouse prototype based on the proposed dynamic indexation scheme and compared its performance with MongoDB [26] (see in [17]).

2.4 Element Calculus

In the frame of our approach to data integration as integration model we consider an advanced XML data model. In fact, data model defines the query language [5]. Based on this, to give declarative queries a new query language (domain element calculus) [14] was developed. A query to XML - database is a formula in element calculus language. To specify formulas a variant of the multisorted first order predicate logic language is used. Notice that element calculus is developed in the style of object calculus [10]. In addition, there is a possibility to give queries by means of λ-expressions. Generally, we can combine the considered variants of queries.

3 Extensible Canonical Data Model

The canonical model kernel is an advanced XML data model: a minor extension of the OPENMath to support the concept of databases. The main difference between our XML data model and analogous XML data models (in particular, XML Schema) is that the concept of data types in our case is interpreted conventionally (set of values, set of operations). More details about the type system of the XML Schema can be found in [3]. A data model concept formalized on the kernel level is referred to as *kernel concept.*

3.1 Kernel Concepts

In the frame of canonical data model we distinguish basic and compound concepts. Formally, a kernel concept is a labeled tree whose leaves are basic kernel concepts. Examples of basic kernel concepts are constants, variables, and symbols (for instance, reserved words). The compound concepts are defined in terms of *binding* and *application* of λ-calculus. The type system is built analogously to that in OPENMath.

3.2 Extension Principle

As we noted above the canonical data model must be extensible. The extension of the canonical model is formed during the consideration of each new data model by adding new concepts to its DDL to define logical data dependencies of the source model in terms of the target model if necessary. Thus, the canonical model extension assumes defining new symbols. The extension result must be equivalent to the source data model. To apply a *symbol* on the canonical model level the following rule has been proposed:

Concept ⟵ *symbol* ContextDefinition

For example, to support the concept of *key* of relational data model, we have expanded the canonical model with the symbol *key*. Let us consider a relational schema example:

S = {S#, Sname, Status, City}

The equivalent definition of this schema by means of extended kernel is considered below:

attribution (S, *type* TypeContext, *constraint* ConstraintContext)
TypeContext ⟵ *application* (*sequence,* ApplicationContext)
ApplicationContext ⟵ *attribution* (S#, *type int*), *attribution* (Sname, *type string*), *attribution* (Status, *type int*), *attribution* (City, *type string*))
ConstraintContext ⟵ *attribution* (name, *key* S#)

It is essential that we use a computationally complete language to define the context [14]. As a result of such

approach, usage of new symbols in the DDL does not lead to any changes in the DDL parser.

3.3 Semantic Level

The canonical model is an XML application. Only syntactical validity of the canonical model concepts representation can be provided on the DTD level. To check semantics the considered application defines new syntactical rules. We define these syntactical rules in content dictionaries.

3.4 Content Dictionaries

The content dictionary is the main formalism to define semantical information about concepts of the canonical data model. In other words, content dictionaries are used to assign formal and informal semantics to all concepts of the canonical data model. A content dictionary is a collection of related symbols, encoded in XML format and fixes the "meaning" of concepts independently of the application. Three kinds of content dictionaries are considered:

- content dictionaries to define basic concepts (symbols);
- content dictionaries to define a signature of basic concepts (mathematical symbols) to check the semantic validity of their representation;
- content dictionary to define a data integration concept.

Supporting the above considered content dictionaries assumes to develop corresponding DTDs. Instances of such DTDs are XML documents. An instance of a DTD of a content dictionary of basic concepts is used to assign formal and informal semantics of those objects. Finally, an instance of a DTD of a content dictionary of a signature of basic concepts contains metainformation about these concepts, and an instance of a DTD of a content dictionary of a data integration concept is a metadata for integrating databases.

3.5 Data Integration Concept

In the frame of our approach to data integration we consider virtual as well as materialized data integration issues within a canonical model. Therefore, we should formalize the concepts of this subject area such as mediator, data warehouse and data cube. We are modelling these concepts by means of the following XML elements: *dbsch*, *med*, *whse* and *cube*.

Mediator. The content of element *dbsch* is based on the kernel *attribution* concept and has an attribute *name*. By means of this concept we can model schemas of databases. The value of attribute *name* is the DB's name. The content of element *med* is based on the elements *msch*, *wrapper*, *constraint* and has an attribute *name*. The value of this attribute is the mediator's name. The element *msch* is interpreted analogously to element *dbsch*. Only note that this element is used during modelling schemas of a mediator. The content of elements *wrapper* and *constraint* is based on the kernel *application* concept. By means of *wrapper* element mappings from source models into a canonical model are defined. The integrity constraints on the level of mediator are the values of the *constraints* elements. It is important that we are using a computationally complete language for defining the mappings and integrity constraints. Below, an example of a mediator for an automobile company database is adduced [5] which is an instance of a content dictionary of data integration concept. It is assumed that the mediator with schema AutosMed = {SerialNo, Model, Color} is integrate two relational sources: Cars = {SerialNo, Model, Color} and Autos = {Serial, Model}, Colors = {Serial, Color}.

```
<cd name = 'dic'>
 <dbsch name = 'Source1'>
  <omattr>
   schema definition of Cars
  </omattr>
 </dbsch>
 <dbsch name = 'Source2'>
  <omattr>
   schema definition of Autos
  </omattr>
  <omattr>
   schema definition of Colors
  </omattr>
 </dbsch>
 <med name = 'Example'>
  <msch>
   <omattr>
    AutosMed: schema  for mediator is  defined
   </omattr>
  </msch>
  <wrapper>
   <oma>
    <oms name = 'convert_to_xml' cd = 'xml'/>
    <oma>
     <oms name = 'union' cd = 'db'/>
     <omv name = 'Cars'/>
     <oma>
      <oms name = 'join' cd = 'db'/>
      <omv name = 'Autos'/>
      <omv name = 'Colors'/>
     </oma>
    </oma>
   </oma>
  </wrapper>
 </med>
</cd>
```

It is essential that, we use a computationally complete language to model the mediator work.

Data warehouse. As we noted above the considered approach to support data warehousing is based on the grid file concept and is interpreted by means of element *whse*. This element is defined as kernel *application* concept and is based on the elements *wsch*, *extractor*, *grid* and has an attribute *name*. The value of this attribute is the name of the data warehouse. The element *wsch* is interpreted in the same way as the element *msch* for the mediator. The element *extractor* is defined as kernel *application* concept and is used to extract data from source databases. The element *grid* is defined as kernel *application* concept and is based on the elements *dim* and *chunk* by which the grid file concept is modelled. To model the concept of stripe of a grid file we introduced an empty element *stripe* which is described by means of five attributes: *ref_to_chunk*, *min_val*, *max_val*, *rec_cnt* and *chunk_cnt*. The values of attributes *ref_to_chunk* are

pointers to chunks crossed by each stripe. By means of *min_val* (lower boundary) and *max_val* (upper boundary) attributes we define "widths" of the stripes. The values of attributes *rec_cnt* and *chunk_cnt* are the total number of records in a stripe and the number of chunks that are crossed by it correspondingly. To model the concept chunk we introduced an element *chunk* which is based on the empty element *avg* and is described by means of four attributes: *id* of type ID, *qty*, *ref_to_db* and *ref_to_chunk*. Values of attributes *ref_to_db* and *ref_to_chunk* are pointers to data blocks and other chunks, correspondingly. Value of attribute *qty* is the number of different points of the considered chunk for fixed dimension. Element *avg* is described by means of two attributes: *value* and *dim*. Values of *value* attributes are used during reorganization of the grid file and contain the average coordinates of points, corresponding to records of the considered chunk, for each dimension. Value of attribute *dim* is the name of the corresponding dimension. To model the concept of dimension of a grid file we introduced an element *dim* which is based on the empty element *stripe* and has a single attribute *name*: i.e. the dimension name.

Data cube. Materialized integration of data assumes the creation of data warehouses. Our approach to create data warehouses is mainly oriented to support data cubes. Using data warehousing technologies in OLAP applications is very important [5]. Firstly, the data warehouse is a necessary tool to organize and centralize corporate information in order to support OLAP queries (source data are often distributed in heterogeneous sources). Secondly, significant is the fact that OLAP queries, which are very complex in nature and involve large amounts of data, require too much time to perform in a traditional transaction processing environment. To model the data cube concept we introduced an element *cube* which is interpreted by means of the following elements: *felement*, *delement*, *fcube*, *rollup*, *mview* and *granularity*. In typical OLAP applications, some collection of data called *fact_table* which represent events or objects of interest are used [5]. Usually, *fact_table* contains several attributes representing dimensions, and one or more dependent attributes that represent properties for the point as a whole. To model the *fact_table* concept we introduced an element *felement* which is based on the kernel *attribution* concept. To model the concept of dimension we introduced an element *delement*. This element is based on the empty element *element* which is described by means of attribute *name*. Value of attribute *name* is the dimension name. The creation of the data cube requires generation of the power set (set of all subset) of the aggregation attributes. To implement the formal data cube concept in literature the CUBE operator is considered [7]. In addition to the CUBE operator in [7] the operator ROLLUP is produced as a special variety of the CUBE operator which produces the additional aggregated information only if they aggregate over a tail of the sequence of grouping attributes. To support these operators we introduced *cube* and *rollup* symbols correspondingly. In this context, it is assumed that all independent attributes are grouping attributes. For some dimensions there are many degrees of granularity that could be chosen for a grouping on that dimension. When the number of choices for grouping along each dimension grows, it becomes non-effective to store the results of aggregating based on all the subsets of groupings. Thus, it becomes reasonable to introduce materialized views.

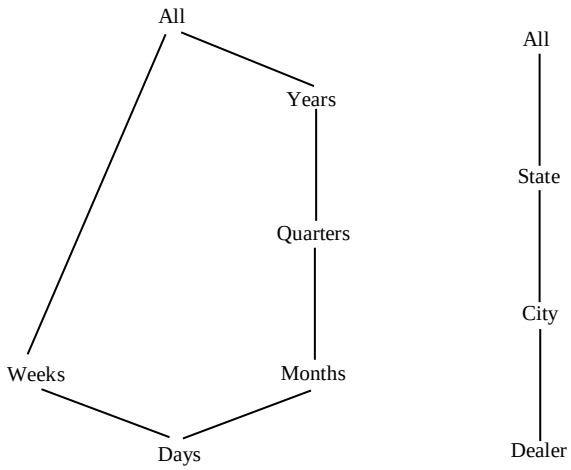


Figure 7 Examples of lattices partitions for time intervals and automobile dealers

Materialized views. A materialized view is the result of some query which is stored in the database, and which does not contain all aggregated values. To model the materialized view concept we introduce an element *mview* which is interpreted by means of an element *view*, and the last is based on the kernel *attribution* concept. When implementing the query over hierarchical dimension, a problem to choose an effective materialized view arises. In other words, if we have aggregated values regarding to granularity Months and Quarters then for aggregation regarding to granularity on Years it will be effective to apply query over materialized view with granularity Quarters. As in [5], we also consider the lattice (a partially ordered set) as a relevant construction to formalize the hierarchical dimension. The lattice nodes correspond to the units of the partitions of a dimension. In general, the set of partitions of a dimension is a partially ordered set. We say that partition P_1 is precedes partition P_2, written $P_1 \leq P_2$ if and only if there is a path from node P_1 to node P_2. Based on the lattices for each dimension we can define a lattice for all the possible materialized views of a data cube which are created by means of grouping according to some partition in each dimension. Let V_1 and V_2 be views, then $V_1 \leq V_2$ if and only if for each dimension of V_1 with partition P_1 and analogous dimension of V_2 with partition P_2 holds $P_1 \leq P_2$. Finally, let V be a view and Q be a query. We can implement this query over the considered view if and only if $V \leq Q$. To model the concept of hierarchical dimension we introduced an element *granularity* which is based on an empty element *partition*, and the latter is described by means of attribute *name*. The value of attribute *name* is the name of the granularity. Below, an example of data cube for an

automobile company database is adduced [5] which is an instance of content dictionary of data integration concept. We consider Sales = {SerialNo, Dealer, Date, Price} as a data cube schema. The considered data cube is implemented on the base of materialized views and is based on three dimensions: Auto, Dealer and Date and has one dependent attribute: Value Set of partitions of dimension Date form a partially ordered set. We are using two granularity elements to represent this set.

```
<cd name = 'dic'>
 ...
 <cube name = 'example'>
  <felement>
   <omattr>
    schema definition of Sales
   </omattr>
  </felement>
  <delement>
   <element name = 'Auto'/>
   <element name = 'Dealer'/>
   <element name = 'Date'/>
  </delement>
  <mview>
   <view name = 'View1'>
    <omattr>
     definition of materialized view Sales1
    </omattr>
   </view>
   <view name = 'View2'>
    <omattr>
     definition of materialized view Sales2
    </omattr>
   </view>
  </mview>
  <granularity name = 'Date'>
   <partition name = 'days'/>
   <partition name = 'months'/>
   <partition name = 'quarters'/>
   <partition name = 'years'/>
  </granularity>
  <granularity name = 'Date'>
   <partition name = 'days'/>
   <partition name = 'weeks'/>
  </granularity>
 </cube>
</cd>
```

The detailed discussion of the issues connected with applying the query language to integrated data is beyond the topic of this paper. Below the XML-formalization of data integration concept is presented.

<!-- include dtd for extended OPENManth objects -->

```
<!ELEMENT cd (dbsch|med|whse|cube)*>
<!ATTLIST cd name CDATA #REQUIRED>
<!ELEMENT dbsch (omattr)+>
<!ATTLIST dbsch name CDATA #REQUIRED>
<!ELEMENT med (msch,wrapper,constraint*)>
<!ELEMENT msch (omattr)>
<!ELEMENT wrapper (oma)>
<!ELEMENT constraint (oma)>
<!ATTLIST med name CDATA #REQUIRED>
<!ELEMENT whse (wsch,extractor,grid)>
<!ELEMENT wsch (omattr)>
<!ELEMENT extractor (oma)>
<!ATTLIST whse name CDATA #REQUIRED>
<!ELEMENT grid (dim+,chunk+)>
<!ELEMENT dim (stripe)+>
<!ELEMENT stripe EMPTY>
<!ELEMENT chunk (avg)+>
<!ELEMENT avg EMPTY>
<!ATTLIST dim name CDATA #REQUIRED>
<!ATTLIST avg value CDATA #IMPLIED
             dim CDATA #REQUIRED>
<!ATTLIST chunk id ID #REQUIRED
        qty CDATA #REQUIRED
        ref_to_db CDATA #REQUIRED
        ref_to_chunk IDREFS #IMPLIED>
<!ATTLIST stripe ref_to_chunk IDREFS #IMPLIED
                 min_val_CDATA #REQUIRED
                 rec_cnt CDATA #REQUIRED
                 max_val_CDATA #REQUIRED
                 chunk_cnt CDATA #REQUIRED>
<!ELEMENT cube (felement,delement,mview?,
                granularity*)>
<!ELEMENT felement (omattr)>
<!ELEMENT delement (element)+>
<!ELEMENT element EMPTY>
<!ATTLIST element name CDATA #REQUIRED>
<!ELEMENT mview (view)+>
<!ELEMENT view (omattr)>
<!ELEMENT granularity (partition)+>
<!ELEMENT partition EMPTY>
<!ATTLIST view name CDATA #REQUIRED>
<!ATTLIST granularity name CDATA #REQUIRED>
<!ATTLIST partition name CDATA #REQUIRED>
```

Figure 8 DTD for formalization of the data integration concept

4 Conclusion

The data integration concept formalization problems were investigated. The outcome of this investigation is a definition language of integrable data, which is based on the formalization of the data integration concept using a mechanism of the content dictionaries of the OPENMath. Supporting the concept of data integration is achieved by the creation of content dictionaries, each of which contains formal definitions of concepts of a specific area of databases.

The data integration concept is represented as a set of XML DTDs which are based on the OPENMath formalism. By means of such DTDs were formalized the basic concepts of database theory, metadata about these concepts and the data integration concept. Within our approach to data integration, an integrated schema is represented as an XML document which is an instance of an XML DTD of the data integration concept. Thus, modelling of the integrated data based on the OPENMath formalism leads to the creation of the corresponding XML DTDs.

By means of the developed content dictionary of the data integration concept we are modelling the mediator and the data warehouse concepts. The considered approach provides virtual and materialized integration of data as well as the possibility to support data cubes with hierarchical dimensions. Within our concept of data cube, the operators CUBE and ROLLUP are implemented. If necessary, in data integrated schemas new super-aggregate operators can be define. We use a computationally complete language to create schemas of integrated data. Applying the query language to the integrated data is generated a reduction problem. Supporting the query language over such data requires additional investigations.

Finally, modern trends of the development of database systems lead to the application of different

divisions of mathematics to data analysis and management. In the frame of our approach to data integration, this leads to the use of corresponding content dictionaries of the OPENMath.

References

[1] Abrial, J.-R.: The B-Book: Assigning programs to meaning. Cambridge University Press, Great Britain, 1996

[2] Briukhov, D. O., Vovchenko A. E., Zakharov V. N., Zhelenkova O. P., Kalinichenko L. A., Martynov D. O., Skvortsov N. A., Stupnikov S. A.: The Middleware Architecture of the Subject Mediators for Problem Solving over a Set of Integrated Heterogeneous Distributed Information Resources in the Hybrid Grid-Infrastructure of Virtual Observatories. Informatics and Applications. Vol. 2, Issue 1, pp. 2-34, Russia, 2008

[3] Date, C. J.: An Introduction to Database Systems. Addison Wesley, USA, 2004

[4] Drawar, M.: OpenMath: An overview. ACM SIGSAM Bulletin, 34(2), 2000

[5] Garcia-Molina H., Ullman J., Widom J.: Database Systems: The Complete Book. Prentice Hall, USA, 2009

[6] Gevorgyan G. R., Manukyan, M. G.: Effective Algorithms to Support Grid Files. RAU Bulletin, N 2, pp. 22-38, Yerevan, 2015

[7] Gray, J., Bosworth, A., Layman, A., Pirahesh, H.: Data Cube: A Relational Aggregation Operator Generalizing Group-By, Cross-Tab, and Sub-Tab. In ICDE, pp. 152-159, USA, 1996

[8] Hindley, J. R., Seldin, J. P.: Introduction to Combinators and λ-Calculus. Cambridge University Press, Great Britain, 1986

[9] Kalinichenko, L. A.: Methods and Tools for Equivalent Data Model Mapping Construction. In EDBT, pp. 92-119, Italy, Springer, 1990

[10] Kalinichenko, L. A.: Integration of Heterogeneous Semistructured Data Models in the Canonical One. In RCDL, pp. 3-15, Russia, 1990

[11] Kalinichenko L. A., Stupnikov S. A. Constructing of Mappings of Heterogeneous Information Models into the Canonical Models of Integrated Information Systems. In Proc. of the 12th East-European Conference, pp. 106-122, Finland, 2008

[12] Kalinichenko L., Stupnikov S. Synthesis of the Canonical Models for Database Integration Preserving Semantics of the Value Inventive Data Models. In Proc. of the 16th East European Conference. LNCS 7503, pp. 223-239, Springer-Verlag, Germany, 2012

[13] Luo, C., Hou, W. C., Wang, C. F., Want H., Yu, X.: Grid File for Efficient Data Cube Storage. Computers and their Applications, pp. 424-429, 2006

[14] Manukyan, M. G.: Extensible Data Model. In ADBIS'08, pp. 42-57, Finland, 2008

[15] Manukyan, M. G., Gevorgyan, G. R.: An Approach to Information Integration Based on the AMN Formalism. In First Workshop on programming the Semantic Web. Available: https://web.archive.org/web/20121226215425/http://www.inf.puc-rio.br/~psw12/program.html, pp. 1- 13, USA, 2012

[16] Manukyan, M. G.: Canonical Data Model: Construction Principles. In iiWAS'14, pp. 320-329, Vietnam, ACM, 2014

[17] Manukyan, M. G., Gevorgyan, G. R.: Canonical Data Model for Data Warehouse. In New Trends in Databases and Information Systems, Communications in Computer and Information Science, Vol. 637, pp. 72-79, Czech Republic, Springer, 2016

[18] Nievergelt, J., Hinterberger, H.: The Grid File: An Adaptable, Symmetric, Multikey File Structure. ACM Transaction on Database Systems, 9(1), pp. 38-71, 1984

[19] Papadopoulos, A. N., Manolopoulos, Y., Theodoridis, Y., Tsoras, V.: Grid File (and family). In Encyclopedia of Database Systems, pp. 1279-1282, 2009

[20] Regnier, M.: Analysis of Grid File Algorithms, BIT, 25(2), pp. 335-358, 1985

[21] Sharma, S., Tim, U. S., Wong, J., Gadia, S., Sharma, S.: A Brief Review on Leading Big Data Models. Data Science Journal, N 13, pp. 138 – 157, 2014. DOI: http/doi.org/10.2481/dsj.14-041

[22] Stupnikov, S. A.: A Varifiable Mapping of a Multidimensional Array Data Model into an Object Data Model, Informatics and Applications. Vol. 7, Issue 3, pp. 22-34, Russia, 2013

[23] Stupnikov, S. A, Vovchenko, A.: Combined Virtual and Materialized Environment for Integration of Large Heterogeneous Data Collections. In Proc. of the RCDL 2014. CEUR Workshop Proceedings 1297:339-348

[24] Stupnikov, S. A, Miloslavskaya, N. G., Budzko, V.: Unification of Graph Data Models for Heterogeneous Security Information Resources' Integration. In Proc. of the International Conference on Open and Big Data OBD 2015 (joint with 3rd International Conference on Future Internet of Things and Cloud, FiCloud 2015). IEEE 2015, Italy, pp. 457- 464, 2015

[25] Zakharov, V. N., Kalinichenko, L.A., Sokolov, I. A., Stupnikov, S. A.: Development of Canonical Information Models for Integrated Information Systems. Informatics and Applications, Vol. 1, Issue 2, pp. 15-38, Russia, 2007.

[26] MongoDB. https://www.mongodb.org

Анализ гуманитарных текстов 1

Text analysis in humanities 1

О подходе к классификации авторефератов диссертаций по темам

© Ю.В. Леонова © А.М. Федотов © О.А. Федотова

Институт вычислительных технологий СО РАН,
Новосибирский государственный университет,
Государственная научно-техническая библиотека СО РАН,
Новосибирск, Россия

juli@ict.nsc.ru fedotov@sbras.ru o4f8@mail.ru

Аннотация. Представлен метод тематической классификации авторефератов диссертаций. Для этого использована специально построенная мера близости документов, учитывающая специфику предметной области. В качестве шкал для определения меры предложено выбирать характеристики структурных атрибутов описания авторефератов (научная новизна; положения, выносимые на защиту, и т. п.). Значения весовых коэффициентов в формуле для вычисления меры близости определяются предполагаемой апостериорной достоверностью данных соответствующей шкалы.

Ключевые слова: мера близости, тематическая классификация, анализ данных, электронные библиотеки.

About Approach to Classification of Thesis Abstracts by Subjects

© Yu.V. Leonova © A.M. Fedotov © O.A. Fedotova

Institute of Computation Technologies of SB RAS,
Novosibirsk State University,
State Scientific and Technical Library of SB RAS,
Novosibirsk, Russia

juli@ict.nsc.ru fedotov@sbras.ru o4f8@mail.ru

Abstract. In this paper the method of thematic classification of the thesis abstracts is considered. It uses specially constructed proximity measure documents, taking into account the specificity of the subject area. The values of the weighting coefficients in the formula for calculating the proximity of the proposed measures are defined a posteriori reliability of the corresponding scale data.

Keywords: analytics proximity measure, thematic classification, analysis of data, electronic libraries.

1 Введение

Поиск и выделение информации являются одной из важнейших задач, возникающих при построении информационных систем. Пользователь ищет не документы как таковые, а сокрытые в них факты или содержимое для удовлетворения собственных информационных потребностей. Универсальным подходом, решающим эту задачу, является тематическая классификация документов. К тому же, как было отмечено Дональдом Кнутом (см. [1]), задачи поиска и классификации документов являются двойственными, поэтому нам достаточно рассмотреть модель классификации документов, наиболее адекватно отражающую особенности работы с информацией.

Труды XIX Международной конференции «Аналитика и управление данными в областях с интенсивным использованием данных» (DAMDID/ RCDL'2017), Москва, Россия, 10–13 октября 2017 года

Наиболее распространенным вариантом классификации библиографических ресурсов является фасетная классификация, теория построения которой формализована индийским библиотековедом Ш.Р. Ранганатаном (см. [2]). Объекты классифицируются одновременно по нескольким независимым друг от друга признакам (фасетам). Применительно к цифровым документам (и электронным ресурсам вообще) в качестве фасетов выступают элементы метаданных, которые включат и ключевые термины. Кратко фасетная классификация состоит в следующем.

Определяется множество тематических классов документов. Класс имеет несколько фасетов, соответствующих различным аспектам классифицируемого понятия.

Из коллекции изучаемых документов выделяются все существенные термины, которые группируются по фасетам, т. е. объединяются в соответствующие классы.

Термин, принадлежащий некоторому фасету, называется его фокусом. При индексировании документов их содержание выражается последовательностью фокусов.

Ниже предложена формальная модель фасетной классификации, основанная на индексации документов ключевыми терминами, выбираемыми из некоторого словаря. Предложен и апробирован алгоритм классификации, основанный на специально построенной мере близости, который учитывает специфику классификационной модели. В качестве базы для экспериментов выбрана коллекция, состоящая из 4000 авторефератов. Мы остановили свой выбор на авторефератах диссертаций по следующим причинам: у них практически одинаковый объем и структура, позволяющие изучить иерархию фасетов.

2 Модель классификации

Простейшая формальная модель классификации документов с использованием метаданных (ключевых терминов) документов выглядит следующим образом [3, 4]. Рассмотрим коллекцию документов $D = \{d_i\}$. Любой документ d_i. из коллекции D представляется как $d_i = <m_i^{j,k}>$, где $m_i^{j,k}$ – значения элементов метаданных T^j, k – количество значений (с учетом повторений) соответствующего элемента метаданных в описании документа. Рассмотрим подмножество метаданных T_C, определяющее набор классификационных признаков документов. Для фиксированного элемента метаданных T^j, где $T^j \subset T_C$, заранее определяются подмножества T_i^j множества значений этого элемента метаданных (указанные подмножества могут, вообще говоря, пересекаться). Множество документов разбивается на классы эквивалентности, соответствующие различным значениям или же заранее выбранным подмножествам множества значений этого элемента метаданных.

Будем считать два документа толерантными, если у них совпадает значение хотя бы одного из элементов метаданных, входящих в T_C или (напомним, что толерантность – отношение, которое обладает свойствами рефлексивности и симметричности, но, вообще говоря, может не обладать, в отличие от отношения эквивалентности, свойством транзитивности). Каждое такое значение порождает класс толерантности [5].

Рассмотрим всевозможные сочетания значений элементов метаданных, входящих в T_C. Множества документов, обладающие одинаковым набором значений, суть ядра толерантности, которые служат классами эквивалентности на множестве документов. С содержательной точки зрения этой ситуации соответствует вхождение некоторого раздела классификатора в раздел более высокого уровня, когда оба этих раздела учитываются при описании пространства толерантности (разумеется, можно и не учитывать раздел более низкого уровня при определении толерантных элементов, но тогда мы будем иметь дело с пространством толерантности, отличным от первоначального).

Таким образом, поисковое предписание, содержащее подмножества метаданных, определяющее набор классификационных признаков и сочетаний значений этих метаданных при помощи логических операций, определяет конкретное ядро толерантности на множестве документов, которое и выдается пользователю в качестве ответа на его информационный запрос. Также на множестве классов толерантности в свою очередь можно ввести отношение толерантности, при этом толерантными считаются классы, имеющие хотя бы один общий документ. Такая конструкция оказывается полезной, например, для организации поиска документов «по аналогии».

3 Мера близости

Предлагаемый подход к построению меры близости (или меры сходства) [6], используемой для классификации документов, основан на понятии толерантности документов, определенной в предыдущем разделе [4].

Ограничимся рассмотрением только ключевых терминов, агрегированных по типам признаков. Количественная характеристика меры близости определяется на множестве документов D следующим образом:

$$m: D \times D \rightarrow [0,1],$$

причем функция m в случае полного сходства принимает значение 1, в случае полного различия – 0. Рассмотрим два документа d_1 и d_2.

Пусть $\boldsymbol{T} = \{\boldsymbol{t_i}\}_{\boldsymbol{i=1}}^{\boldsymbol{M}}$ – упорядоченный (каким-либо образом, например, лексиграфически) список ключевых терминов, входящих в оба документа, с учетом повторений (где M – общее количество ключевых терминов). Вычисление меры близости осуществляется по следующей формуле:

$$m(d_1, d_2) = \sum_{i=1}^{M} a_i \, m_i(d_1, d_2),$$

где i – номер элемента метаданных (ключевого термина), $m_i(d_1, d_2)$ – мера близости по i элементу (иными словами по i шкале), α_i – весовые коэффициенты. Поскольку в описываемой ситуации практически все шкалы – номинальные (состоящие из дискретных текстовых значений), то мера сходства по i-й шкале определяется следующим образом: если значения i-х элементов документов совпадают, то мера близости равна 1, иначе 0.

Весовые коэффициенты должны удовлетворять следующим условиям: $\sum_{i=1}^{M} \alpha_i = 1$, $\alpha_i = \alpha_j$, если значение термина t_i совпадает со значением термина t_j.

Пусть $\boldsymbol{P} = \{\boldsymbol{p}_k\}_{\boldsymbol{k=1}}^{\boldsymbol{N}} -$ список уникальных ключевых терминов, входящих в оба документа, M_k –

число повторений термина p_k. Тогда меру близости можно переписать так:

$$m(d_1, d_2) = \sum_{k=1}^{N} (a_k * M_k)(m_k / M_k),$$

α_k – весовой коэффициент, соответствующий значению термина p_k, m_k – число совпадений термина p_k в документах d_1 и d_2. Мы получаем новые весовые коэффициенты $\beta_k = \alpha_k M_k$, которые уже характеризуют конкретный ключевой термин. Нетрудно видеть, что

$$\sum_{k=1}^{N} \beta_k = 1.$$

Отметим, что мы здесь автоматически получаем весовой коэффициент, который пропорционален частоте встречаемости термина. Кроме того, при задании меры можно принять во внимание тот факт, что значения весовых коэффициентов β_k определяются предполагаемой апостериорной достоверностью данных соответствующей шкалы, и в определённых случаях один из коэффициентов может быть увеличен с пропорциональным уменьшением остальных. Например, полное (или даже «почти полное») совпадение значений какого-либо атрибута документов d_1 и d_2 может быть более весомо в случае, когда количество значений этого атрибута в документе d_1 достаточно велико (по сравнению со случаем, когда документ d_1 имеет всего одно совпадение).

4 Построение схемы взвешивания

Документ представляется в виде множества ключевых терминов (термов). Множество всех термов $T = \{t_0, \ldots, t_{|T|}\}$. Каждому терму $t_i \in T$ сопоставляется некоторый вес ω_{ij}, $0 \le \omega_{ij} \le 1$, характеристика (действительное число) встречаемости слова в документе $d_j \in D$.

Учет разделов. Документ может содержать разделы, имеющее разную значимость, с точки зрения вклада в тематическое сравнение двух документов, например, в автореферате «положения, выносимые на защиту» и «апробация работы». Для учета веса каждого раздела в меру близости добавляются весовые коэффициенты:

$$m(d_1, d_2) = \sum_{r=1}^{R} \psi_r \, m_r(d_1, d_2),$$

где R – число разделов, ψ_r – весовой коэффициент, учитывающий априорную значимость r-го раздела, $0 \le \psi_r \le 1$, $m_r(d_1, d_2)$ – мера близости по r-му разделу.

Рассмотрим наиболее популярную схему взвешивания TF-IDF.

Схема взвешивания TF (term frequency – частота терма). Каждому термину, встретившемуся в документе, присваивается вес, который зависит от количества появлений этого термина в данном документе. Таким образом, оценивается важность терма t_i в пределах отдельного документа d_j.

Пусть fr_{ij} – число вхождений терма t_i в документ d_j. Тогда частота терма определяется как

$$TF(t_i, d_j) = \frac{fr_{ij}}{\sum_i fr_{ij}}, \text{ где } 0 \le i \le |T|,\ 0 \le j \le |D|.$$

Отметим, что эта характеристика уже присутствует при построении меры близости.

Схема взвешивания IDF (inverse document frequency – обратная частота документа) – инверсия частоты, с которой термин встречается в документах коллекции. Учёт IDF уменьшает вес широкоупотребительных слов. Для каждого уникального термина в пределах конкретной коллекции документов существует только одно значение IDF.

$$IDF(t_i, D) = \frac{|D|}{|(d_i \subset t_i)|},$$

где $|D|$ – количество документов в коллекции, $|(d_i \supset t_i)|$ – количество документов, в которых встречается t_i (когда $fr_i \neq 0$), $0 \le i \le |T|$.

Априорный вес некоторого термина пропорционален количеству употребления этого термина в документе и обратно пропорционален частоте употребления термина в других документах коллекции. Кроме того, априорный вес термина зависит от экспертной оценки его значимости.

Выбор метода классификации

Разработав качественный алгоритм классификации, можно создать автоматический классификатор текстов. Алгоритм классификации текстов работает по принципу: текстовый документ, состоящий из выделенных терминов, при заданных классах классифицируется в класс, который имеет максимальное значение меры близости документа к классу с учетом весовых коэффициентов разделов автореферата и терминов. Для решения задачи классификации документов разработано много методов, обеспечивающих точность классификации, сравнимую с выполненной человеком-экспертом.

5 Практические результаты

В качестве исходных данных для тестирования алгоритма классификации использовались авторефераты диссертаций по 4 темам: «распознавание образов» (тема 1), «распознавание речи» (тема 2), «геоинформационные системы» (тема 3), «онтологии, описание предметной области» (тема 4). В эталонные наборы для каждой тематики вошли по 30 авторефератов.

Формирование списка ключевых терминов (словаря) является отдельной задачей [4]. Например, словарь ключевых терминов может формироваться экспертом на основе его знаний о предметной области. В нашем случае список сформирован на основе текстов эталонных авторефератов, его объем составил 192 ключевых слова.

Таблица 1 Точность и полнота классификации «своих» и «чужих» документов

Раздел	Точность						Полнота					
	Свои			Чужие			Свои			Чужие		
	По разделам	С учетом многозначности	По тексту	По разделам	С учетом многозначности	По тексту	По разделам	С учетом многозначности	По тексту	По разделам	С учетом многозначности	По тексту
Тема1	0,75	0,807	0,685	1	1	1	0,831	0,865	0,747	0,214	0,5	0,25
Тема2	0,647	0,647	0,47	0,833	0,833	1	0,647	0,647	0,477	0,833	0,83333	0,8
Тема3	0,771	0,9	0,5	1	1	1	0,843	0,915	0,579	0,333	0,88889	0,083
Тема4	0,651	0,647	0,578	0,666	0,642	0,625	0,686	0,656	0,637	0,533	0,6	0,333

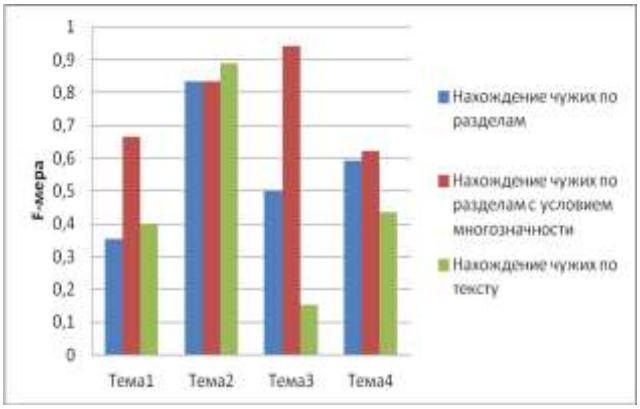


Рисунок 1 F-мера при нахождении «чужих» документов

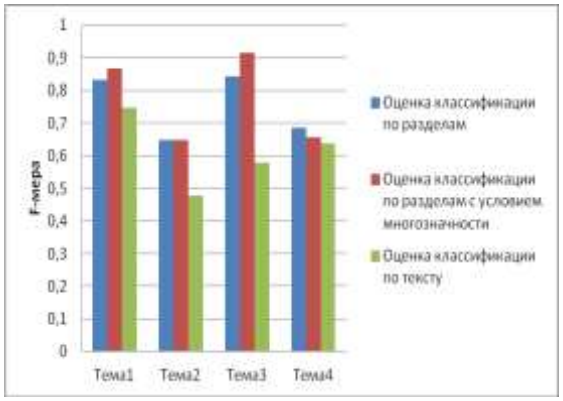


Рисунок 2 F-мера нахождения и классификации «своих» документов

Классификация проводилась по следующему алгоритму. Первоначально для каждой темы на основе эталонного набора находился центроид – характерный набор ключевых терминов с весами, который потом использовался для сравнения. Далее вычислялась мера близости проверяемого автореферата к центроиду класса (темы).

Отметим, что наиболее удовлетворяет приведенным выше требованиям алгоритм Роше, использующий меру близости и весовые коэффициенты. Класс характеризуется эталонным представителем – документом (возможно, искусственным), содержащим в себе все усредненные признаки класса. Таким образом, центроид является эталонным представителем класса. Каждому признаку экспертом приписывается вес, отражающий его важность. При классификации определяется минимальное расстояние между эталонными и классифицируемым документами.

6 Результаты тестирования

На вход системы было подано по 4000 ранее неизвестных текстов авторефератов. Для классификации использовался весь текст автореферата, из которого выделялись значимые ключевые слова. Мера близости рассчитывалась по выделенным ключевым терминам в словаре для каждой темы. При классификации информация выделялась из следующих разделов автореферата: Актуальность темы исследования, Цели и задачи; научная новизна, Объект и предмет исследования, Теоретическая и практическая значимость работы, Методология и методы исследования, Положения, выносимые на защиту, Степень достоверности и апробация результатов.

Для каждого раздела вычислялась мера близости, итоговая мера близости вычислялась как среднее значение.

Тестирование алгоритма проводилось в трех режимах: 1) классификация по разделам автореферата; 2) классификация по разделам с проверкой многозначности терминов – если термин принадлежит нескольким темам, то проверяется тематическая принадлежность соседних терминов; 3) классификация по всему тексту автореферата без выделения разделов.

Принадлежность автореферата теме определяется по превышению порога близости между тестируемым авторефератом и центроидом темы.

В таблице 1 и на рисунках 1, 2 представлены результаты вычисления точности, полноты и F-меры, полученные при тестировании алгоритма. Видно, что наилучшей точностью нахождения чужих документов обладает метод поиска по тексту автореферата, однако полнота и F-мера – наихудшие. Лучшими характеристиками обладает метод поиска по разделам с условием многозначности. Наихудший параметр точности соответствует Теме 3 – геоинформационные системы, что обусловлено присутствием некоторых терминов, как «пространственное распределение»,

«пространственная структура» и т. п., в текстах химической направленности. Дополнение словаря химическими терминами позволит повысить точность классификации.

Заключение

На основании анализа полученных данных можно сделать следующие выводы. Алгоритм классификации по всему тексту автореферата дает неплохие результаты в случае, когда надо отсеивать «чужие» документы. На практике обычно это и требуется в большинстве случаев. Однако в случае, когда известно, что у каждого документа есть тема, он проигрывает двум другим алгоритмам, выполняющим классификации по разделам автореферата. Алгоритм классификации по разделам с условием многозначности терминов показывает себя не хуже в поиске чужих документов, как алгоритма классификации по тексту, так и алгоритма классификации по разделам. В тестах алгоритм классификации по всему тексту автореферата немного превосходит алгоритм классификации по разделам с условием многозначности терминов. Однако алгоритм классификации по разделам с условием многозначности терминов вырывается вперед по сравнению с другими алгоритмами при поиске «своих» документов.

Литература

[1] Кнут, Д.: Искусство программирования. Том 1. Основные алгоритмы. The Art of Computer Programming. Vol. 1. Fundamental Algorithms / под ред. Ю.В. Козаченко. Москва: Вильямс, 720 с. (2002)

[2] Ранганатан, Ш.Р.: Классификация двоеточием. Основная классификация / пер. с англ. М.: ГПНТБ СССР (1970)

[3] Федотов, А.М., Барахнин, В.Б.: Проблемы поиска информации: история и технологии. Вестник НГУ. Серия: Информационные технологии, 7 (2), сс. 3-17 (2009)

[4] Федотов, А.М., Барахнин, В.Б., Жижимов, О.Л., Федотова, О.А.: Модель информационной системы для поддержки научно-педагогической деятельности. Вестник Новосибирского гос. ун-та. Серия: Информационные технологии, 12 (1), сс. 89-101 (2014)

[5] Шрейдер, Ю.А.: Равенство, сходство, порядок. М.: Наука (1971)

[6] Воронин, Ю.А.: Начала теории сходства. Новосибирск: Наука. Сиб. отд-е, 128 с. (1991)

Применение инструментов интеллектуального анализа текстов в юриспруденции

© Д.С. Зуев ©А.А. Марченко ©А.Ф. Хасьянов

Казанский (Приволжский) федеральный университет,
Казань, Россия

dzuev11@gmail.com anton.marchenko@kpfu.ru ak@it.kfu.ru

Аннотация. Описана архитектура системы интеллектуального анализа текстов в юриспруденции, способной на имеющейся базе данных судебных документов выявлять общие зависимости, предоставлять для ознакомления юридические дела, близкие по тематике, рекомендовать наиболее вероятные исходы судебного рассмотрения или помечать важные места, на которые следует обращать внимание при процессуальных действиях с использованием инструментов текстовой аналитики.

Ключевые слова: аналитика и управление данными, интенсивное использование данных, электронные библиотеки, кластеризация, рекомендательная система, микросервисная архитектура.

Text Mining Tools in Legal Documents

© D.S. Zuev ©A.A. Marchenko ©A.F. Khasiannov

Volga Region Federal University,
Kazan, Russia

dzuev11@gmail.com anton.marchenko@kpfu.ru ak@it.kfu.ru

Abstract. We present the architecture of the system for the intellectual textual analysis in jurisprudence based on microservices. The system can identify common dependencies on an existing database of legal documents, provide legal cases close to each other, familiarize them with the most probable outcomes of judicial review or mark out important places during procedural actions.

Keywords: analytics and data management, data intensive domains, digital libraries, clustering, recommender system, microservices.

1 Введение

Как известно, информационное общество характеризуется высоким уровнем развития информационно-коммуникационных технологий (ИКТ) и их интенсивным использованием всеми и всюду. В основе ИКТ лежит информация, а сами они во многом определяют содержание, масштабы и темпы развития других технологий.

Интересным направлением разработки специализированных автоматизированных информационных систем является создание интеллектуальных систем, способных на имеющейся базе данных судебных документов выявлять общие зависимости, предоставлять судьям для ознакомления близкие по тематике дела, рекомендовать наиболее вероятные исходы или помечать важные места, на которые судебным работникам следует обращать внимание при процессуальных действиях. Подобная система, на наш взгляд, поможет участникам судебного процесса точнее оценивать свои позиции или выбирать лучшую стратегию поведения, а судьям – в сжатые сроки формировать подборку связанных документов, не тратя для вынесения вердикта лишнего времени на поиск во всем архиве документов.

Проведенные исследования по семантическому структурированию информации в других предметных областях (см., например, [1, 2]), анализ инструментов текстовой аналитики (см, например, [3]) и наработки по применению семантических технологий при работе с юридическими документами [4] говорят о реализуемости поставленной задачи.

2 Интеллектуальная система «Робот-юрист»

2.1 Цели и задачи

«Робот-юрист» – это информационная система, которая должна позволять участникам юридического процесса правильно проводить подготовку дела, а также осуществлять планирование судебной деятельности. Эта система ориентирована на арбитражные суды, занимающиеся рассмотрением споров в сфере предпринимательства. В целом наш проект направлен на развитие российского правового государства, обеспечение доступности,

Труды XIX Международной конференции «Аналитика и управление данными в областях с интенсивным использованием данных» (DAMDID/ RCDL'2017), Москва, Россия, 10–13 октября 2017 года

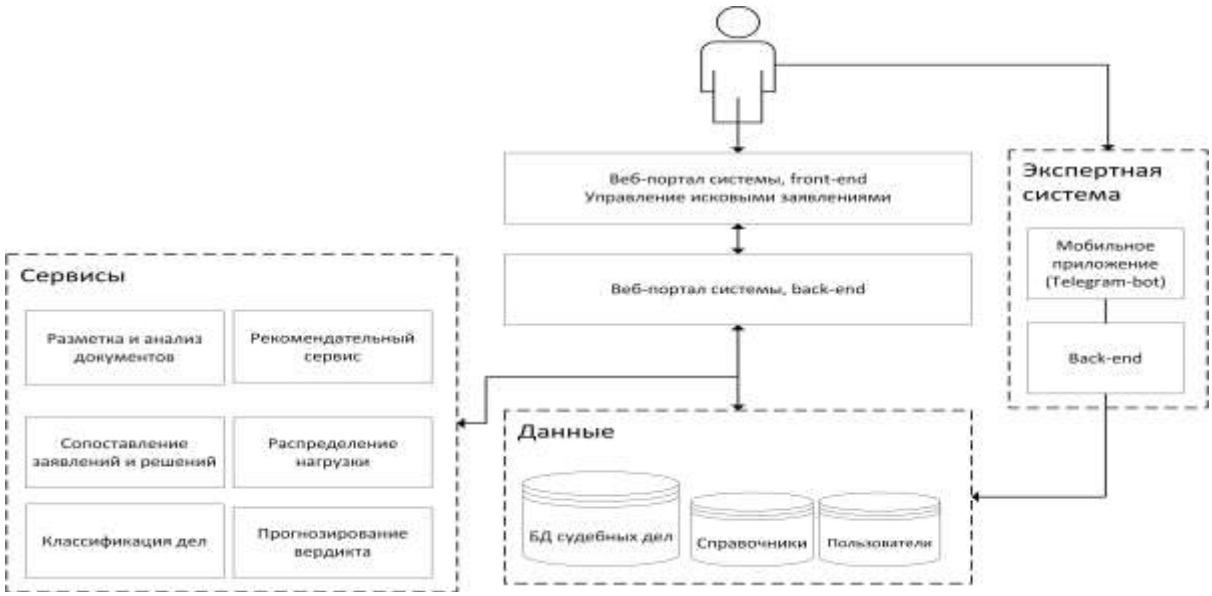


Рисунок 1 Архитектура системы

открытости и прозрачности правосудия, формирование у граждан правосознания, основанного на верховенстве Права.

Задача системы – помочь определить характер спора, осуществить поиск и проверку действия правовых норм, регулирующих спорные правоотношения, оказывать содействие в установлении компетентного суда (подсудность, подведомственность), статуса участников спора (действующее, ликвидированное, банкрот), определении круга обстоятельств, имеющих значение для рассмотрения спора, характера спорного правоотношения, нормы права, подлежащей применению (действует ли данная норма), а также проверять достаточность и комплектность представляемых документов. Отдельными функциями планируются обеспечение возможности оформления искового заявления, а также вычисление (по предоставленным исходным данным на основе архива судебных дел) вероятности принятия того либо иного решения.

Для достижения поставленных целей были поставлены следующие задачи: создание портала для формирования шаблонов исковых заявлений с отслеживанием их жизненного цикла; разметка и анализ существующей базы судебных решений, исковых заявлений (классификация заявлений и решений, извлечение сущностей и фактов); подбор аналогичных дел и решений, рекомендательный сервис; сопоставление исковых заявлений и судебных решений; распределение судебных дел между судьями с учетом их специализации и текущей загрузки и прогнозирование вероятного решения по предоставленным исходным данным.

Каждая из выделенных задач является автономным модулем разрабатываемой информационной системы, а сама система – практическая демонстрация совместного использования ряда семантических технологий и инструментов текстовой аналитики.

2.2 Архитектура системы

Текущие парадигмы разработки предусматривают два концептуально различных подхода к дизайну приложений. Первый вариант – «монолитные приложения», когда вся логика по обработке запросов выполняется в рамках единственного процесса, при этом используются возможности конкретного языка программирования для разделения приложения на классы и функции. Однако любые изменения, даже самые небольшие, требуют перекомпиляции всего дистрибутива информационной системы и последующего обновления всех ее модулей. С течением времени изменения в логике работы одного модуля начинают влиять на функции других модулей.

Другой подход – это построение среды, в которой отдается предпочтение слабым связям, абстрагированию низкоуровневой логики, гибкости, а также возможности многократного использования и обнаружения компонентов [5, 6], сервис-ориентированная архитектура (Service-Oriented Architecture, SOA). Такая архитектура строится на сервисах, а не на приложениях. Сервисы – это дискретные программные компоненты, предоставляющие четко определенную функциональность и используемые в составе многих приложений. Каждый сервис представляет собой изолированную сущность с минимумом зависимостей от других совместно используемых ресурсов. Таким образом, возникает возможность изменять отдельные сервисы, не затрагивая при этом всю систему. Дальнейшим развитием парадигмы сервис-ориентированной архитектуры можно считать появление архитектуры микросервисов [7]. Термин «Microservice Architecture» получил распространение в последние несколько лет для описания способа проектирования приложений в

виде набора независимо развертываемых сервисов.

Архитектурный стиль микросервисов – это подход, при котором единое приложение строится как набор небольших сервисов, каждый из которых работает в рамках собственного процесса и взаимодействует с остальными. Сервисы построены вокруг бизнес-потребностей и развертываются независимо с использованием полностью автоматизированной среды. Централизованное управление минимизировано, а сами сервисы могут быть написаны на разных языках программирования и использовать разные технологии хранения данных. Более того, внутри каждого микросервиса вполне может быть задействована собственная база данных (см. [7]).

С учетом достаточно большого количества модулей системы наиболее логичным путем для создания «Робота-юриста» стало применение архитектуры микросервисов.

Архитектура системы приведена на Рис. 1. Нами выделено несколько групп сервисов, взаимодействующих между собой с помощью программного интерфейса (API). Каждый из них реализует одну из соответствующих функциональных задач. На схеме выделены серверная и клиентская часть веб-портала системы, а также слой доступа к данным – база данных судебных дел и решений, нормативно-справочная информация. В виде отдельного модуля разрабатывается экспертная система, в автоматическом режиме оказывающая консультации по юридическим вопросам в формате взаимодействия с виртуальным собеседником – Telegram-Ботом.

2.3 Разметка массива документов

Разметка существующего массива документов необходима для дальнейшего обучения сервисов системы. Для реализации этой задачи использовался инструмент для быстрого структурированного аннотирования текстов BRAT [8]. BRAT – это веб-система с открытым исходным кодом, разработанная группой разработчиков в университетах Токио и Манчестера. Результаты разметки получаются в виде, удобном для дальнейшей машинной программной обработки.

Судебные решения и дела открыты и доступны для просмотра в интернете и представляют собой массив неразмеченных документов, в котором ориентироваться непросто. Важна собственно разметка текстов судебных дел для выделения классов и подклассов сущностей, их зависимостей с целью дальнейшего построения модели машинного обучения.

На текущий момент времени, в рамках создания прототипа системы, принято решение о первоначальной разметке сравнительно небольшого количества документов (около 3000). Важно отметить, что самих типов споров, значит, и классов связанных документов может быть достаточно много. С целью упрощения работы на начальном этапе мы обрабатывали судебные дела, относящиеся только к нескольким категориям судебных споров.

Размеченный текст будет использоваться для обучения подсистемы поиска аналогов и прогнозирования вердикта по делу. В качестве результата работы получаем размеченный текст, который записывается в БД судебных дел для дальнейшей обработки.

Для первоначальной разметки были выделены основные сущности, такие, как «Истец», «Ответчик», «Предмет спора», «Действующие нормы». На текущий момент времени определено 56 сущностей, которые необходимо выделять внутри судебных решений для дальнейшей обработки. Множество выделенных сущностей будет уточняться по мере увеличения объема размеченного текста. На сегодняшний день проведена всего лишь первая итерация данного процесса.

2.4 Рекомендательный сервис

Одной из важнейших задач формируемой информационной системы являются поиск и предоставление аналогичных решений по схожим судебным искам. Таким образом, необходим сервис поиска аналогичных документов, или рекомендательный сервис.

Существуют два основных типа рекомендательных систем: контент-ориентированные и социальные (коллаборативной фильтрации) (см., например, [9]). Первые основаны на представлении предпочтений пользователей путем анализа содержимого рекомендательных элементов. Системы второго типа моделируют предпочтения, оценивая близость профилей пользователей. Ниже под рекомендательным сервисом будем понимать информационную систему, которая: 1) формирует модель предметной области на основе массива документов (включая подготовительные операции – приведение к векторному виду, кластеризацию и т. п.); 2) получает на вход документ и выдает список документов, близких к входному.

По сравнению с поисковыми системами рекомендательные системы наиболее полезны, когда у пользователя возникают трудности с формулировкой эффективного поискового запроса.

Подходы к организации рекомендательных сервисов могут быть разными, в [1] описан подход с использованием онтологий и предпочтений пользователей. Учитывая специфику предметной области и разрабатываемой системы, использовать предпочтения пользователей не корректно.

Алгоритм работы сервиса можно разделить на два этапа. На подготовительном этапе обрабатываются все имеющиеся документы: вырезаются знаки пунктуации, термы приводятся к единому виду (для слов с разными окончаниями и суффиксами). Далее документ приводится к векторному виду. Для представления массива документов в виде числовых векторов, отражающих важность использования каждого слова из некоторого набора слов

(количество слов набора определяет размерность вектора), в каждом документе используется мера TF-IDF [3, 10]. На основе массива векторов происходит кластеризация.

На первом шаге необходимо определить количество *K* кластеров, мы использовали для этого формулу $K = N_{doc}/10$, где N_{doc} – общее количество обрабатываемых документов. Далее производится собственно кластерный анализ по методу *K-means* (метод К-средних, [3, 11]). Полученные результаты сохраняются для дальнейшего использования.

На основном этапе работы на вход сервису подается идентификатор документа. Производится приведение его к векторной форме, которая обрабатывается моделью, причисляется к определенному кластеру. На выходе алгоритм выдает первые *n* документов из того же кластера, что и входной документ, количество выдаваемых документов настраивается, на данном этапе реализации системы *n* определено равным 10.

Процесс переобучения модели следует проводить периодически, например, раз в сутки, либо после существенного изменения всего корпуса документов.

Обработка массива из 3250 документов занимает 5 мин (Intel® Core™ i7-3632QM CPU @ 2.20GHz × 8), что на текущем этапе развития системы «Робот-юрист» является приемлемым показателем быстродействия. Сервис реализован на языке Python, взаимодействие с другими модулями системы происходит по внутреннему согласованному протоколу взаимодействия.

2.5 Классификация судебных дел

Одной из проблем судебного делопроизводства является процедура определения категории и характера спора. Правильное определение категории судебного спора важно, поскольку влияет на назначение судьи на соответствующий процесс, а назначаемый судья должен иметь максимальный опыт рассмотрения подобных споров. На текущий момент выявлено около 60 различных категорий судебных споров, которые встречаются с разной частотой. За определение категории судебного дела отвечает модуль классификации судебных дел. Процесс классификации с ростом количества обрабатываемых документов может быть очень затратным по времени, поэтому с архитектурной точки зрения было решено вынести данную функциональность как отдельный микросервис с реализацией обмена с другими модулями системы в асинхронном режиме. К тому же определение категории спора (судебного дела) не является задачей, требующей мгновенного ответа.

На уровне межсервисного взаимодействия общий алгоритм обработки документа выглядит следующим образом: на вход подается идентификатор документа; из документа выделяются ключевые слова и их количество; проводятся анализ и подбор класса дела; алгоритм возвращает идентификатор класса судебного дела, который становится дополнительным свойством документа. При добавлении нового класса проводятся анализ допустимых ключевых слов и повторное обучение нейронной сети.

К сожалению, на текущий момент нами окончательно не выбран оптимальный способ реализации данной задачи – рассматривается реализация алгоритма с использованием глубинного обучения и сверточных нейронных сетей или с использованием латентно-семантического анализа.

2.6 Создание шаблонов исковых заявлений

Отдельной задачей является сопоставление судебных актов и заявлений по рассмотренным делам, поскольку сами исковые заявления, в отличие от базы знаний принятых решений, являются закрытыми и не публикуются в сети интернет. В рамках разработки системы «Робот-юрист» актуальной является задача связывания вновь поданного искового заявления и близких результатов судебных процессов для дальнейшей обработки. В этом случае необходимо иметь заявление в размеченном виде, удобном для машинной обработки. Для этого необходимо либо отдельно предусматривать процесс разметки массива электронных копий бумажных исковых заявлений, либо формировать заявления изначально в электронном виде и далее распечатывать готовое заявление с помощью системы. Второй вариант является предпочтительным, и его было предложено реализовать в рамках создания прототипа системы.

Для получения экземпляров исковых заявлений сразу в электронном виде был предложен механизм веб-портала – шаблонизатора заявлений. При подаче пользователем системы искового заявления система формирует печатную версию заявления в соответствии с регламентирующими нормативными документами РФ, а электронная копия документа автоматически размечается и сохраняется в базе данных системы с определенным статусом.

Процесс организован следующим образом: пользователь авторизуется на портале системы; ему предоставляется ряд экранных форм с полями ввода для заполнения данных. После окончания ввода данных пользователь сохраняет заявление в системе; в базе данных системы появляется размеченный вариант документа для дальнейшего анализа, а пользователю предоставляется печатная форма заполненного искового заявления.

Веб-портал предусматривает несколько ролей пользователей с различной функциональностью, также предложена и реализована статусная модель судебного дела для удобства отслеживания жизненного цикла документа в системе.

2.7 Экспертная система

В рамках проекта также разрабатывается решение по автоматизации предоставления экспертных консультаций по вопросам юридического характера. Решение представляет собой экспертную систему (ЭС) (см., например, [12]) – компьютерную систему, способную частично заменить эксперта-специалиста

в разрешении какой-либо проблемы юридического характера.

В рамках проекта реализована экспертная система в области защиты интеллектуальной собственности. Важными вопросами в автоматизации предоставления экспертных консультаций являются надежность решений и удобство использования, поэтому решения ЭС подкрепляются ссылками на соответствующие нормативные документы, указанные юристами при формировании базы знаний.

Были определены наиболее часто встречающиеся сценарии и вопросы в данной области права. На текущий момент реализованы 13 типовых сценариев поведения ЭС, которые практически полностью покрывают всевозможные случаи в данной области права.

В качестве пользовательского интерфейса к экспертной системе был выбран интерфейс чат-бота или, другими словами, виртуального собеседника, реализованного в виде Telegram-Бота (далее – бота). Совпадение логики процессов взаимодействия с ботом и ЭС позволяет предоставить удобный доступ к инструментам юридического консультирования со всех платформ, для которых доступен сам мессенджер (Telegram). Логика работы модуля представляет собой конечный автомат, а использование бота в качестве интерфейса к ЭС позволяет снизить трудозатраты на разработку пользовательского интерфейса и сконцентрироваться на функционале ЭС вследствие простоты разработки.

3 Заключение

Теоретические исследования в рамках текстовой аналитики показывают наличие готовых или практически готовых инструментов для реализации функций отдельных модулей системы. Необходимы лишь их грамотное объединение и применение в отдельно взятых предметных областях. «Робот-юрист» должен стать именно такой демонстрацией применения известных подходов и алгоритмов в юриспруденции.

На данный момент завершен первый этап создания системы – закончено проектирование системы и реализован прототип системы «Робот-юрист», производится разметка документов. Для успешного завершения работ и перевода в опытную эксплуатацию требуется дальнейшая оптимизация как различных алгоритмов текстовой аналитики, так и пользовательского интерфейса. Выбранная архитектура построения приложения позволяет производить модификацию отдельных модулей системы, не затрагивая общего механизма взаимодействия. Также необходимы апробация инструментов системы на большем массиве документов и рефакторинг программного кода.

Поддержка

Работа выполнена за счет средств субсидии, выделенной Казанскому федеральному университету для выполнения государственного задания в сфере научной деятельности, проект 2.8712.2017/БЧ.

Литература

[1] Елизаров, А.М., Жижченко, А.Б. Жильцов, Н.Г., Кириллович А.В., Липачёв, Е.К.: Онтологии математического знания и рекомендательная система для коллекций физико-математических документов. Докл. Академии наук, 467 (4), сс. 392-395 (2016). doi: 10.1134/S1064562416020174

[2] Елизаров, А.М., Липачёв, Е.К., Невзорова О.А., Соловьев, В.Д.: Методы и средства семантического структурирования электронных математических документов. Докл. Академии наук, 457 (6), сс. 642-645 (2014). doi 10.7868/S0869565214240049

[3] Ингерсолл, Грант С., Мортон, Томас С., Фэррис, Эндрю Л.: Обработка неструктурированных текстов. Поиск, организация и манипулирование / Пер. с англ. Слинкин А.А. М.: ДМК Пресс, 414 с.: ил. (2015)

[4] Peroni, S.: SemanticWeb Technologies and Legal Scholarly Publishing Law, Springer, Governance and Technology Series, 15 (2014). doi 10.1007/978-3-319-04777-5

[5] Gold, N. et al.: Understanding Service Oriented Software. IEEE Software, 21 (2), pp. 71-77 (2004)

[6] Jones, S.: Toward an Acceptable Definition of Service. IEEE Software, 22 (3), pp. 87-93 (2005)

[7] Fowler, M.: Microservices a definition of this new architectural term. https://martinfowler.com/articles/microservices.html

[8] Stenetorp, P., Pyysalo, S., Topić, G., Ohta, T., Ananiadou, S., Tsujii, J.: Brat: a Web-based Tool for NLP-Assisted Text Annotation. Proc. of the Demonstrations Session at EACL (2012)

[9] Ricci, F., Rokach, L., Shapira, B., Kantor, P.B.: Recommender Systems Handbook. N.Y.: Springer (2011)

[10] https://ru.wikipedia.org/wiki/TF-IDF

[11] https://ru.wikipedia.org/wiki/K-means

Джарратано, Дж., Райли, Г.: Экспертные системы. Принципы разработки и программирование. 4-е издание. Вильямс, 1152 с. (2007)

Double Funding of Scientific Projects: Similarity and Plagiarism Detection

© Denis Zubarev[1,2] © Ilya Sochenkov[1,2] © Ilya Tikhomirov[1] © Oleg Grigoriev[1]

[1] Federal Research Center Computer Science and Control of the Russian Academy of Sciences, Moscow, Russia

[2] Peoples' Friendship University of Russia (RUDN University), Moscow, Russia

zubarev@isa.ru sochenkov_iv@rudn.university tih@isa.ru grigoriev@isa.ru

Abstract. The paper addresses the problem of scientific projects double funding. New methods for text similarity and plagiarism detection in scientific projects are proposed. An empirical evaluation of proposed methods is given. The data was provided by the Russian Foundation for Basic Research.

Keywords: text similarity, plagiarism detection, scientific projects, double funding.

1 Introduction

The problem of double funding of scientific projects is well-known and, unfortunately, not solved yet. It's a common problem for the whole scientific community. For example, neuroscientist Steven McIntire of the University of California submitted a five-year, US$ 1.6-million grant application to the US National Institutes of Health (NIH) in November 2001. But five months earlier, the US Army had awarded him $1.2 million for a project with strikingly similar scientific aims. Both grants were given for investigation of genes that affect responses to ethanol in Caenorhabditis elegans (roundworm), which is used as a model organism to understand the effects of alcohol in humans [1]. This case is not unique - researchers often submit proposals to various organizations and often win several grants for very similar investigations.

In the paper, we propose double funding detection method. It is based on text similarity and plagiarism detection techniques. The method is tested on applications forms for young scientists' research grants of the Russian Foundation for Basic Research (RFBR).

2 Method for double funding detection

We provide short description of our method in this section. We will not go in much detail, because of the limit of the paper's length. The more detailed description can be found in the paper [4].

The classic information retrieval approach is used - there are two processing stages: data indexing and plagiarism retrieval. The retrieval stage contains two steps: search for similar projects (projects on the same topic), and matching similar text fragments.

2.1 Scientific projects indexing

At the first step we index previously extracted text information from grant applications and scientific reports. One of the main research goals is to detect semantically identical text parts, such as rephrasing (further called as "reused text"). That's why we perform deep linguistic analysis, which includes part-of-speech tagging, syntactic parsing, semantic role labeling, and semantic relation extraction [2,3]. Due to the Russian language structure the above information helps to detect semantically identical or similar text parts.

Linguistic analysis results are stored in originally designed index subsystem that allows efficiently store and search heterogeneous semantic networks [4]. Additionally we build a spectral inverted index [5]. This index maps single words and stable word phrases in normal form to their logarithmic TF (LTF) weights. The index doesn't contain all stable word phrases. Different heuristics (frequency, IDF-weight) are used to make a decision about a word phrase inclusion.

Index subsystems support incremental update, so new grant applications and reports can be indexed shortly after they are uploaded. After the database is filled out, search for similar projects can be performed as the first step of the retrieval stage.

2.2 Search for similar projects

We suppose that significant text reuse is inherent for documents belonging to the same or similar topic. Topics are not predefined – they are described implicitly by the most top-weighted (TF-IDF is used for weighting) single words (in normal form) and word phrases (also in normal form). IDF weights are calculated based on word and phrase frequencies of all indexed texts. We employ vector space model (VSM) for searching for documents with similar topics. We use vector of relatively small size (100-200 elements), each element is a TF-IDF weight of a corresponding word or phrase. The index subsystem (more precisely the spectral inverted index) provides vectors of indexed documents that are overlapped with the query vector by K%. For calculating distance between vectors the asymmetric Hamming is employed [6]. Finally, we choose only documents with similarity score higher than the predefined threshold. Also there is

Proceedings of the XIX International Conference "Data Analytics and Management in Data Intensive Domains" (DAMDID/RCDL'2017), Moscow, Russia, October 10–13, 2017

a limit by the number of similar documents which will be propagated to the next step.

2.3 Matching similar text fragments

On this step, we perform sentence-wise matching. Before that, we skip too short sentences, which are less than five words, and sentences that contain mostly non-alphanumeric symbols (for example, large formulas). Then all sentences from the query document are compared with all sentences from documents chosen on the previous step. A sentence is represented as a list of identifiers of normalized words. Since most of pairs of sentences are not similar, they are filtered out. We use the fast algorithm to estimate the size of the intersection of two sets [7]. If two sentences share at least 50% of words, we measure their similarity score. We score two sentences by comparing various characteristics of words with the same normal form (further called as common words).

- Word overlap measure: The value of the measure is the normalized sum of IDF-weights of common words between two sentences.
- Syntactic similarity measure: The value of this measure is the number of common words with the same syntactic relations in their sentences. The value is normalized by the total number of common words.
- Semantic similarity measures: There are two semantic measures, one for comparison of semantic roles and one for comparison semantic relations of common words [2, 3]. The values of these measures are the number of common words with the same semantic roles and the number of common words that are in the same semantic relations. The values are normalized by the total number of common words.

The final score is a weighted sum of the described above measures. The weight for each measure is chosen during parameters optimization [4]. The matching by lexis has the significant impact on the overall score. But the syntactic/semantic measures allow to represent a sentence not as a bag-of-words but as a coherent text with a meaning. A value of these measures will be significantly lower for sentences with the same set of words but with different semantic usage of these words.

2.4 Post-processing

We consider a sentence to be a reused one if the similarity score exceeds a predefined threshold. All sentences that are considered as reused ones are merged in larger text fragments if they are close to each other in the text. Then we calculate the percentage of the reused sentences to the total amount of sentences in the text. If this percentage is larger than X% the project is marked as suspicious. The final decision is made by the experts, who explore the plagiarism detection results when reviewing grant applications and reports.

2.5 Method evaluation

We performed an evaluation of our method on special corpus for plagiarism detection. The evaluation is described in the paper [4]. Also we participated in PAN 2014 where we obtained the second highest F-measure for the source retrieval track [8].

3 Experiments

3.1 Dataset description

The Russian Foundation for Basic Research kindly provided us with information on grant applications submitted for programs to support young scientists: ”mol_a” 2016, “mol_ev_a” 2016, “mol_a_mos” 2015, “mol_a_dk” 2016. “Mol_a” is a program for small teams of young researchers (under 35 years old). “Mol_a” covers wide range of research areas: biology, computer science, mathematics and physics, etc. The task of “mol_ev_a” is to provide young scientists with the opportunity to conduct a scientific research on topics relevant to the development of critical technologies and high-tech products. The set of allowed scientific topics is fixed and rather small. “Mol_a_mos” is a program for young PhDs that is partly funded by the government of Moscow. It has a specific set of topics, which is focused on improving life in modern metropolises (e.g. ecology, new materials, telecommunication etc.). “Mol_a_dk” aims to support young PhDs in their studies. The set of topics is similar to “mol_a”. “Mol_a_dk” and “mol_a_mos” are personal grants. All these programs were started at the same time and all applications were submitted approximately at the same time period. Each form contains project name, short abstract and some other fields. The text fields are presented as a plain text (see Table1), dataset size is presented in Table 2.

Table 1 Textual fields of the application forms

Field name	Description
Name	Name of the project
Abstract	Short abstract
State	Description of the State-of-the-art in the research area
Methods	Proposed methods and approaches
Goal	The fundamental goal of the research
Results	Expected scientific results
Prev_results	Previous scientific results
Publications	List of authors publications
Plan	General work plan

Table 2 Dataset size

Research program	Application form count
mol_a	7361
mol_ev_a	701
mol_a_mos	288
mol_a_dk	1621
Total	**9971**

Unfortunately, we cannot publish the dataset, because we depend on the data provider terms of usage. Therefore it is not possible to compare our method with other

plagiarism detections tools, since it requires uploading texts to the external detector. But it is forbidden by the terms of usage.

3.2 Experiment setup

The goal of the experiment was to detect similarity between sets of application forms, described above. Results should be represented as similar projects sets. This information was very useful for the experts: very similar projects are not eligible for RFBR funding. Grant applications were preprocessed and loaded to the index subsystem that was described in the previous section. While indexing, the subsystem stores special tag that indicates a field of the form, to which indexed text belongs. But we didn't use this specific information further, when reused text was searched. It simplified our task as we considered each application as a plain text. After indexing, we applied described in the previous section method to each grant application. In the end we obtained pairs of documents with the percentage rate of reused text.

There were two rounds of comparison. Firstly, all applications from the "mol_a" and "mol_ev_a" were compared against each other. We performed search against all indexed documents, but applications from other, than "mol_a" and "mol_ev_a" programs were skipped. We picked pairs, with percent of reused text greater or equal to 30%. Secondly, we performed the same search procedure for all grant applications from "mol_a_mos" and "mol_a_dk" programs. At this step each pair of applications with 10 percent or greater of reused text was chosen as suspicious. The choice of collections and minimal thresholds were proposed by the experts from RFBR. Such a choice can be explained that those programs are very similar and were started nearly at the same time. The main difference between them is the different set of allowed scientific topics. The participants may have been very tempted after submitting to the program with fixed set of topics ("mol_a_mos", "mol_ev_a"), to try luck in another program ("mol_a_dk", "mol_a" respectively), where was no restriction on topics. Also experts didn't want to deal with a large amount of false positives that would be caused by searching in all collections of projects. The results of comparison were taken into consideration by RFBR experts.

3.3 Experiment results

The percentage distribution of reused text is shown on the next picture. This is the distribution for the first round of comparison: "mol_a_mos" and "mol_a_dk" to each other.

It shows that most applications reused little amount of text from other applications. It is worth noting that most pairs with percentage greater than 50% (6 from 7 pairs) were from the same authors. However, applications were submitted to different programs. This pinpoints one strategy of getting double funding: submitting the same or slightly modified applications to the different programs that are launched at the same time. The pie-chart with combinations of all pairs is presented in the next figure.

The distribution of the percentage of a reused text for the second round of comparison: "mol_a" and "mol_ev_a", is shown in the Figure 3.

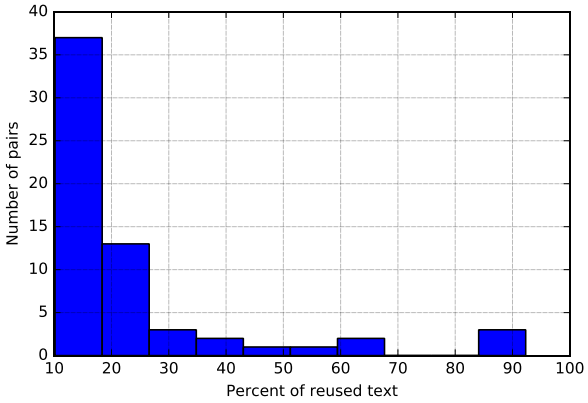


Figure 1 Distribution of the reused text percentage for "mol_a_mos" and "mol_a_dk" programs applications

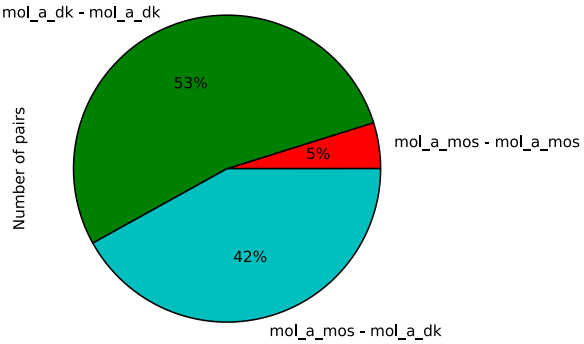


Figure 2 All combinations of pairs of applications with reused texts

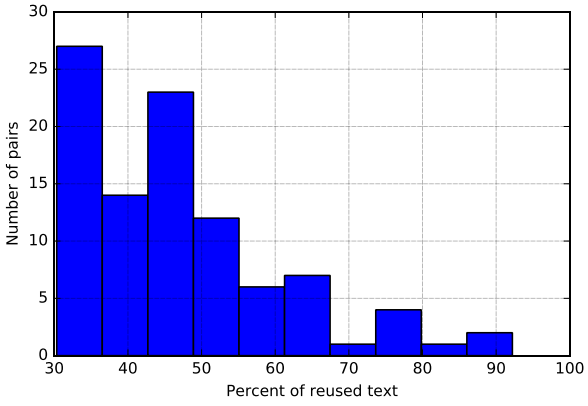


Figure 3 Distribution of the reused text percentage for "mol_a" and "mol_ev_a" programs applications

There are more pairs (28) of applications that have more than 50% of shared text. And most of these pairs (79%) are from the same program. This program is "mol_a" except one pair from "mol_ev_a". There are multiple cases when text was reused in three and even six applications. These applications use the same "template" and only the subject of an application is different. It seems that this strategy of getting double funding is based on "spamming" the same applications, in hope that

some of them may be funded twice. This strategy is only observed in programs with relatively high amount of applications: mol_a, mol_a_dk. The figures 2 and 4 show that there are substantially less projects with reused text for programs with smaller amount of applications (mol_a_mos and mol_ev_a programs).

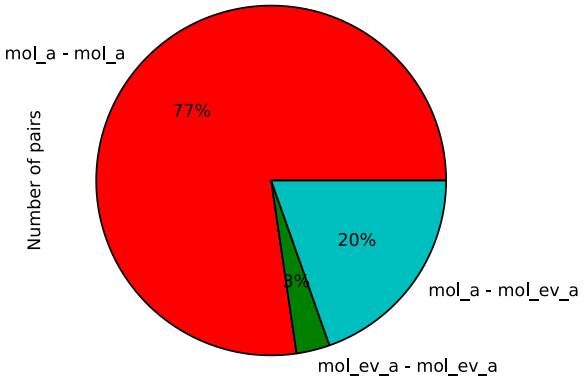


Figure 4 All combinations of pairs of applications with reused texts

4 Conclusion and future work

In the paper we present the method for double funding detection and its empirical evaluation on the dataset of the Russian Foundation for Basic Research. Our future work will be devoted to the development of a large-scale system for double funding detection for different funding organisations. We plan to develop more fine-grained comparison of applications. It will perform grouping of similar fields of application forms into one text and compare only related groups of text (e.g. group of ”The fundamental goal of the research” and “Expected scientific results” will be compared against ”Previous scientific results” and so on). Also we will focus on visualizing results of comparison and will work on a tool that may help experts to analyze the pair of applications with high rate of shared text. It will highlight reused text and when clicking on it will provide context of source text with links to the section of its origin.

Support

The research is supported by Russian Foundation for Basic Research, project 16-29-12881.

References

[1] Reich, E. S., & Myhrvold, C. L.: Funding agencies urged to check for duplicate grants. Nature, 493, pp. 588–589 (2013)

[2] Shelmanov A. O., Smirnov I. V: Methods for semantic role labeling of Russian texts. Computational Linguistics and Intellectual Technologies. Papers from the Annual International Conference "Dialogue" No. 13, pp. 607–620 (2014)

[3] Osipov, G., Smirnov, I., Tikhomirov, I., Shelmanov, A.: Relational–situational method for intelligent search and analysis of scientific publications. In: Proceedings of the Workshop on Integrating IR technologies for Professional Search, in conjunction with the 35th European Conference on Information Retrieval. Volume 968, pp.57–64 (2013)

[4] Zubarev D. V., Sochenkov I. V.: Paraphrased Plagiarism Detection Using Sentence Similarity. Computational Linguistics and Intellectual Technologies. Papers from the Annual International Conference "Dialogue" 2017, v. 1, pp. 408–418 (2017)

[5] Shvets, A., Devyatkin, D., Sochenkov, I., Tikhomirov, I., Popov, K., Yarygin, K.: Detection of current research directions based on full-text clustering. In: Proceedings of Science and Information Conference, IEEE pp. 483–488 (2015)

[6] Suvorov R. E., Sochenkov I. V.: Establishing the similarity of scientific and technical documents based on thematic significance. Scientific and Technical Information Processing, vol. 42(5), pp. 321-327 (2015)

[7] Takuma, D., Yanagisawa, H.: Faster upper bounding of intersection sizes. In: Proceedings of the 36th international ACM SIGIR conference on Research and development in information retrieval, pp.703–712 (2013)

[8] Zubarev D., Sochenkov I.: Using Sentence Similarity Measure for Plagiarism Source Retrieval, In CLEF (Working Notes), pp. 1027-1034 (2014)

Проекты анализа данных в различных ОИИД

Data analysis projects in various DID

Standardization of Storage and Retrieval of Semi-structured Thermophysical Data in JSON-documents Associated with the Ontology

© A.O. Erkimbaev © V.Yu. Zitserman © G.A. Kobzev © A.V. Kosinov

Joint Institute for High Temperatures, Russian Academy of Sciences, Moscow, Russia

adilbek@ihed.ras.ru vz1941@mail.ru gkbz@mail.ru kosinov@gmail.com

Abstract. A new technology for data management of a complex and irregular structure is proposed. Such a data structure is typical for the representation of the thermophysical properties of substances. This technology based on storage of data in JSON files is associated with ontologies for the semantic integration of heterogeneous sources. Advantages of JSON-format – the ability to store data and metadata within a text document, accessible perceptions of a person and a computer and support for the hierarchical structures needed to represent semi-structured data. Availability of a multitude of heterogeneous data from a variety of sources justifies the use of the Apache Spark toolkit. When searching, it is supposed to combine SPARQL and SQL queries. The first one (addressed to the ontology repository) provides the user with the ability to view and search for adequate and related concepts. The second, accessed by JSON documents, retrieves the required data from the body of the document. The technology allows to overcome a variety of schemes, types and formats of data from different sources and implement a permanent adjustment of the created infrastructure to emerging objects and concepts not provided for at the stage of creation.

Keywords: thermophysical properties, semi-structured data, JSON-format, ontology.

1 Introduction

The constantly increasing volume and complexity of the data structure on the substances and materials properties imposes stringent requirements for the information environment that integrates diverse resources belonging to different organizations and states. In contrast to the earth science or medicine, here the source of data is the growing publication flow. In so doing the volume of data is determined not so much by the number of objects studied, as by the unlimited variety of conditions for synthesis, measurement, morphological and microstructural features, and so on. It can be said that of the three defining dimensions of Big Data (the so-called "3V-Volume, Velocity, Variety" [3]), it is the latter plays a decisive role, that is, an infinite variety of data types.

In this paper, we propose a set of solutions borrowed from Big Data technology, allowing to overcome with minimum expenses two main difficulties in the way of integration of resources. The first one is the variety of accepted schemes, terminologies, types and formats of data and so on, and the second is the need for permanent adaptation of the created structure to the emerging variations in the nomenclature of terms (objects, concepts etc) not provided at the design stage. The need for variation in the data structure can be associated with the expansion of the range of substances (e.g. by including nanomaterials), the range of properties (e.g. by including state diagrams), or by changing the data type, say with the transition from constants to functions.

The solutions proposed in the work are based on the joint use of previously used technologies:

- Data interchange standard in the form of text-based structured documents, each of which is treated as an atomic storage unit;
- Ontology-based data management;
- General framework **Apache Spark** for large-scale data processing.

The three main elements of the planned data infrastructure (Figure 1): a plethora of primary data sources of different structures (databases, file archives, etc.) subject to conversion to the standard JSON format; ontologies and controlled dictionaries for the semantic integration of disparate data; Big Data technology for storing, searching and analyzing data.

2 Data preparation

2.1 General scenario

The general scheme of data preparation (Figure 2) assumes as an initial material a large body of external resources, thematically related, but arbitrary in terms of volume, structure and location. Among them are sources of structured data which include factual SQL databases (DB), document-oriented DB, originally structured files in **ThermoML** [7] or **MatML** [9] standards, numerical tables in CSV or XLS formats and so on. The second group (possibly dominant in terms of volume) is formed by unstructured data: text, images, raw experimental or modeling data etc.

Proceedings of the XIX International Conference "Data Analytics and Management in Data Intensive Domains" (DAMDID/RCDL'2017), Moscow, Russia, October 10–13, 2017

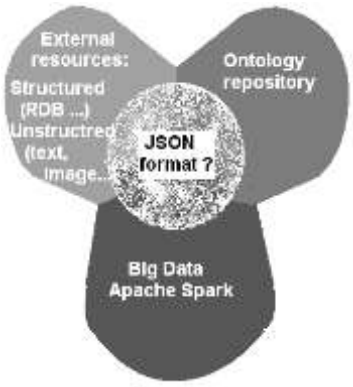


Figure 1 Key elements of the data infrastructure concept

The first stage of data preparation is the downloading of records from external sources with their subsequent conversion to the standard form of JSON documents [2]. In so doing, the conversion of structured documents can be entrusted to software whereas the unstructured part is subject to "manual" processing with the extraction of relevant information from the texts. Finally, the control element in this scheme is the repository of subject specific (domain) and auxiliary ontologies.

The distinctive characteristic of the proposed approach is that the starting data sources remain "isolated" and unchanged. Resource owners periodically download data to JSON files by templates linked with ontological models.

In so doing they determine themselves the composition, amount and relevance for the "external" world of the data being download. This type of interaction is passive, in contrast to active, when client can use the **JDBC** or **ODBC** interface to access databases.

2.2 JSON-documents

The basic unit of storage is a structured text document recorded in JSON format, one of the most convenient for data and metadata interchange propose [2]. The advantage of JSON-document – text-based language-independent format, is easy to understand and quickly mastered, a convenient form of storing and exchanging arbitrary structured information. Previously, structured text based on the XML format was proposed as a means of storing and exchange thermophysical data in the **ThermoML** project [7] and data on the properties of structural materials in the **MatML** project [9].

Here, a text document is proposed as the main storage unit, written in JSON format, which is less overloaded with details, simplifying the presentation of the data structure, reducing their size and processing time. In particular, the JSON format is shorter, faster read and written, can be parsed by a standard JavaScript function, rather than a special parser, as in the case of the XML format [https://www.w3schools.com/js/js_json_xml.asp].

Among other advantages of the format, one can note a simple syntax, compatibility with almost all programming languages, as well as the ability to record hierarchical, that is, unlimited nested structures such as "key-value". By way of **value** may be accepted object (unordered set of key-value pairs), **array** (ordered set of values), string, number (integer, float), **literal** (true, false, null). It is also important that the JSON format is a working object for some platforms, in particular for **Apache Spark**, allowing for the exchange, storage and queries for distributed data.

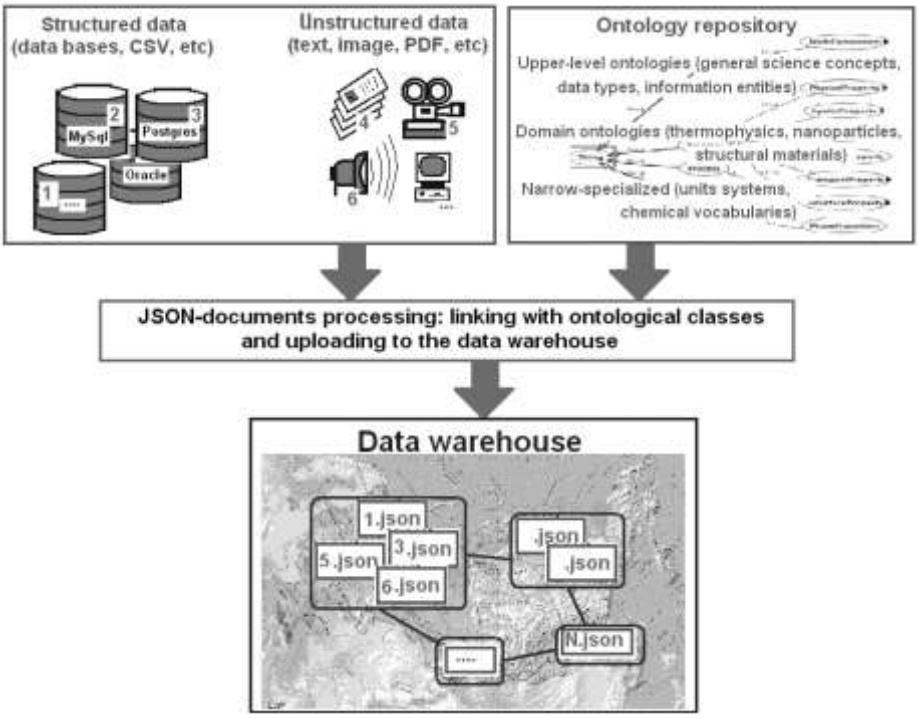


Figure 2 Schematic sketch of initial data processing

The rich possibilities of JSON-format as a means of materials properties data interchange attracted the attention of developers of the Citrination platform [10]. They proposed JSON-based hierarchical scheme **PIF** (physical information file), detailing the object, **System**, whose fields include three data groups, explaining what an object is (name, ID), how it was created/synthesized the generality of the created scheme should be sufficient for storing objects of arbitrary complexity, "from parts in a car down to a single monomer in a polymer matrix". Flexibility of the **PIF**-scheme is achieved due to additional fields and objects, as well as the introduction of the concept of **category**. This concept is nothing but a version of the scheme, oriented to a certain kind of objects, say substances with a given chemical identity.

2.3 Ontology-based data management

The second stage of data preparation is the linking of extracted metadata with concepts from ontologies and dictionaries assembled into a single repository. The management of the repository is entrusted to an **ontology-based data manager**, which allows for the search and editing of terms (class) ontologies, as well as their binding to JSON documents, Figure 3. This means that when the particular source schema is converted to a JSON format, terms from ontologies, rather than source attributes, are used as its keys. It is also possible to use additional keys for a detailed description of the data source itself, for example, indicating the type of DBMS, name and format of textual or graphical file, authorship and other official data, "sewn up" in the atomic "unit" of storage.

The role of ontologies is to introduce semantics (a common interpretation of meaning) into documents, as well as the ability to adjust the data structure of the JSON-documents by editing the ontology.

Linking documents with ontologies allows to perform semantic *(meaningful*) data search (more precisely, metadata) using SPARQL queries, which makes it possible to reveal the information of the upper and lower levels (*super and sub-classes*) and side-links (*related terms*), without knowing the schema of the source data. Thus, the user can view and retrieve information without being familiar with the conceptual schema of a particular DB or the metadata extracted from unstructured sources.

The repository should include three types of ontologies and controlled vocabularies: *upper-level*, *domain* and *narrow-specialized*. The first type is scientific top-level ontology, which introduces the basic terminology used in different fields, for example such concepts as **substance, molecule, property, state**, as well as informational entities that reflect the representation of data: **data set, data item, document, table, image,** etc. Most of these terms and links between them can be borrowed from ontologies presented on the server **Ontobee** [11], for example **SIO** (Semanticscience Integrated ontology) or **CHEMINF** (Chemical Information ontology).The second type of ontology (domain ontology) should cover the terminology of certain subject areas, for example, thermophysics, structural materials, nanostructures, etc. For each of the domains, as a rule, some ontologies previously created and presented in publications or Web are already available on the basis of which it is possible to build and further maintain its own subject-specific ontology. Finally, the third type (narrow-specialized) should include ontologies or vocabularies for systems of units (for example, **UO** on the above portal **Ontobee**) or chemical dictionaries, for example **ChemSpider** [6] and the like. Figure 3 illustrates the binding of terms from a JSON document to ontological terms.

Even at the stage of data preparation the proposed technology provides:

- consistency with accepted standards regardless of the structure and format of the original data;
- semantic integration of created JSON-documents;
- inclusion of previously not provided objects and concepts by expanding classes or introducing new ontologies and dictionaries.

The scheme of the generated data is determined by the initial data scheme with subsequent correction in the process of linking with the terms and structure of the corresponding ontological model. It should be noted that JSON-documents are objects with which one can operate using external API. In so doing, there is always the possibility of accessing keys in JSON documents not currently linked with a particular ontology term.

At the same time, it seems justified to identify or bind not only keys, but also values with ontological terms. For example, the key/value pair "Property": "Heat Capacity" is presented in Figure 3. This will allow in the future to facilitate the formation of SQL query, relying on the information received from the ontologies repository.

The experience of using ontology in the data interchange through text documents has already been implemented in a special format **CIF** (*Crystallographic Information File*) [8], intended for crystallographic information.

In other cases of using a JSON document for the storage of scientific data (for example, in the mentioned Citrination system [10]), the categorization and introduction of new concepts is carried out by a special commission without linking with the concepts of ontological models.

3 Technique of storage and access to data

Given the increasing volume and distributed nature of the data on the properties, some of the Big Data technologies would be appropriate for infrastructure design. Their advantage is due not so much to high performance in parallel computing, but rather to a pronounced orientation to work with data (storage, processing, analysis and so on) in a distributed environment (when data sources are located on remote severs). Among the available open-source means, the **Apache Spark** high-performance computing platform [5] is offered here. Along with other technological features, it is distinguished by the presence of built-in libraries for complex analytics including running SQL-queries. By means of SQL-queries one can access the contents of structured JSON documents. It is the ability of SQL-queries to data plays a key role in the task of their

integration. The efficiency of Spark in the storage and processing of data is also associated with its ability to maintain interaction with a variety of store types: from **HDFS** (Hadoop Distributed Files System) to traditional database on local servers. We should also note the built-in library **GraphX** – an application for processing graphs, which provides our project with our own tools for working with ontologies.

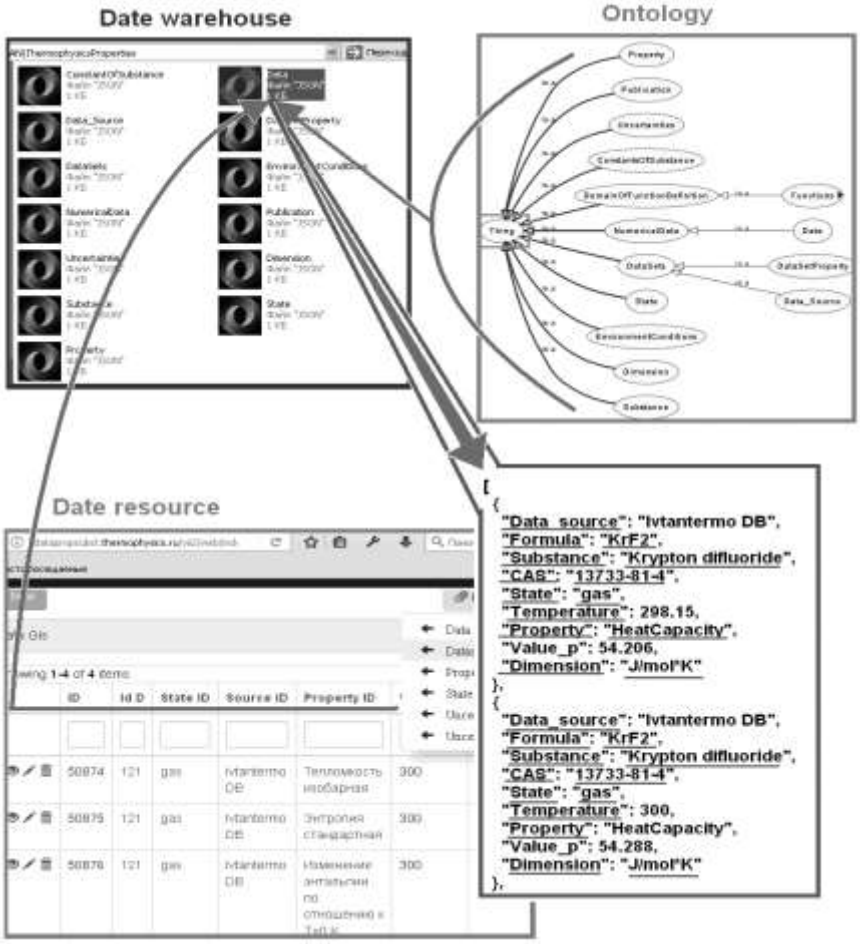


Figure 3 Linking JSON documents to ontology classes using the example of the ontology for the domain of thermophysical properties

The computing platform shown in Fig. 4 includes three basic elements:

• Management-dispatching of distributed computing under the control of Hadoop YARN, Apache Mesos or stand-alone;

• Computer interface – API for languages Scala, Java и Python;

• Powerful and diverse means of data storage.

Advantages of **Apache Sp**ark in comparison with other technologies (MapReduce) – higher computing speed and the ability to handle data of different nature (text, semi-structured and structured data, etc.) from various sources (files of different formats, DBMS and streaming data). It is also important to have APIs for Scala, Java and Python and high-level operators to facilitate code writing, integration with the **Hadoop** ecosystem [4], which unites libraries and utilities provided for in Big Data technology. The ecosystem has I/O interfaces (InputFormat, OutputFormat) that are supported by a variety of file formats and storage systems (**S3, HDFS, Cassandra, Hbase, Elasticsearch** and so on).

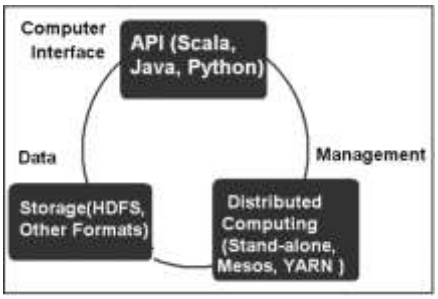


Figure 4 Spark Architecture

For storage purposes, the Apache Spark provides for interaction with three groups of data sources:

Files and file systems – local and distributed file systems (**NFS, HDFS, Amazon S3**), capable of storing data in different formats: text, JSON, SequenceFiles (binary key/value pairs) etc;

Sources of structured data available through **Spark SQL**, including JSON, **Apache Hive**;

Relational databases and key/value pairs storages (**Cassandra, HBase**), accessed by built-in and external libraries of interaction with the databases such as **JDBC/ODBC** or search engine **Elasticsearch**.

In doing so **Spark** supports loading and saving data from different formats files: unstructured (text), semistructured (JSON), structured (such as CSV or SequenceFiles). The **Apache Spark** operation reduces to the formation and transformation of **RDD** (*Resilient Distributed Dataset*) sets, which are distributed collections of elements. When parallel processing, the data is automatically distributed in sets between the computers in the cluster. **RDD** sets can be created by importing **HDFS** files using the Hadoop InputFormats tool or by converting from another **RDD**.

The main task (access and search among structured and semistructured data) is implemented by the **Spark SQL** module, which is one of the built-in Apache Spark libraries. **Spark SQL** defines a special type of **RDD** called **SchemaRDD** (in recent versions, the term **DataFrame** is used). The **SchemaRDD** class represents a set of objects row, each of which is a normal record. The **SchemaRDD** type is called a schema (a list of data fields) of records. **SchemaRDD** supports a number of operations that are not available in other sets, in particular, execution of SQL-queries. **SchemaRDD** sets can be created from external sources (JSON-documents, Hive tables), query results and from ordinary **RDD** sets.

Three main features of **Spark SQL**:

- It can download data from different sources;
- It requests data using SQL within Spark programs and from external tools related to **Spark SQL** through standard mechanisms for connecting to databases via **JDBC/ODBC**;
- It provides an interface between SQL and ordinary code (when used within Spark SQL programs), including the ability to connect sets of **RDD**s and SQL tables.

It is possible to configure **Spark SQL** to work with structured data **Apache Hive**. The **Apache Hive** store is specifically designed for querying and analyzing large data sets, both inside **HDFS** and in other storage systems. **JDBC/ODBC** standards are also supported by **Spark SQL**, which allows executing a direct SQL query to external relational databases, in the case of the above defined active type of interaction.

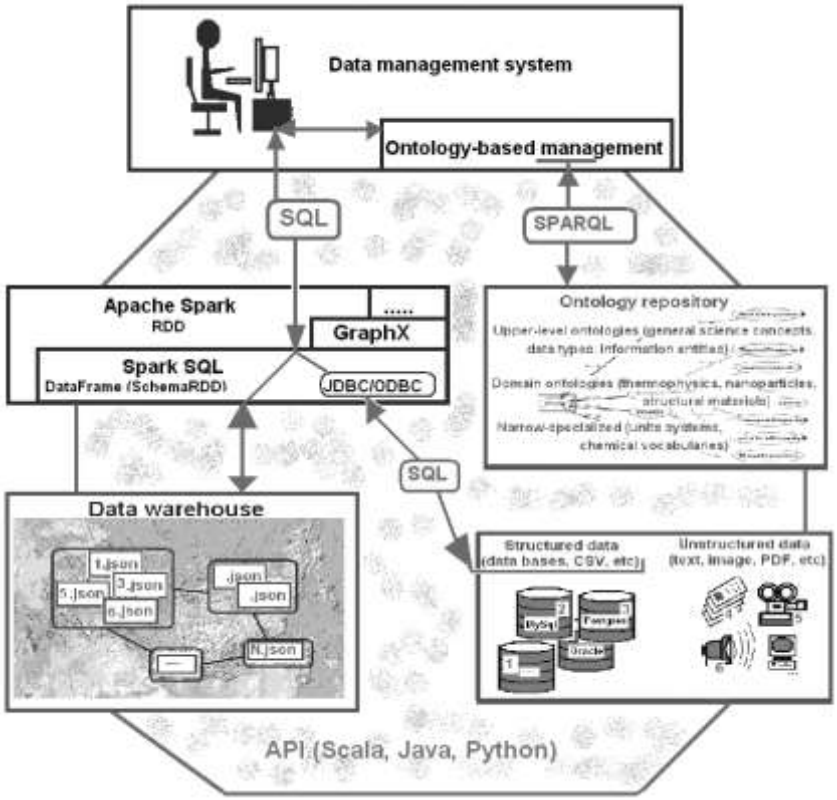


Figure 5 Web-environment for managing heterogeneous and distributed data on the substance properties (databases, unstructured and semistructured files and so on)

The main scenario that uses the **Apache Spark** features is shown in Figure 5. As a result of uploading data to JSON documents according to the above procedure, we will have data sets with a single classifying key system identical to terms from ontologies. The organization of data requests in JSON documents will always be based on definitions from this single system. Thus, one can form the SQL query of interest in the interface of the data processing system. In this case, it always remains possible to access the ontology repository to refine or supplement the terms of the query using the **SPARQL** query (Figure 5). Then the SQL-request coming from the user interface initiates the work of the Spark SQL module. As a result of the work of Spark SQL module, **RDD** or **DataFrame** sets are created, including the selected records, which can be processed by the system's service functions for further use. In fact, the user's work consists of two phases: viewing ontologies terms with the choice of adequate for the formation of SQL-query; access to the repository of

processed data with a SQL query. Thus, the main scenario involves unifying heterogeneous data by converting them to JSON documents and processing them using **Spark SQL**. Other scenarios are also justified, if we take into account the diversity of the source data. For example, the **Spark SQL** module allows direct query to relational databases without their conversion to the JSON format. On the other hand, you can provide access to JSON documents by collecting them in a file system using other Big Data tools. The first and main feature of JSON data collection based on ontological models terms – the unambiguous interpretation of the content and type of data. In this case users and external programs can freely work with data, because the ontology term, mapped to a key or value in the body of the JSON file, has available and accepted definitions and properties. For example, links to various types of files (graphics, multimedia, exe-files, etc.) can be described adequately and functionally using keys-terms from ontologies describing data formats. The second feature, as it is not strange, is the possibility of including in the exchange of such data sources that do not allow active access or changes due to various reasons. Then uploading the data to an external JSON file solves this problem, providing independent data storage and their full description via ontological models.

The listed technologies, supported by **Apache Spark**, provide unlimited productivity and variety of opportunities to handle complex data, which include data on the properties of substances, including traditional materials and nanostructures.

References

[1] Ataeva, O.M., Erkimbaev, A.O., Zitserman, V.Yu. et al.: Ontological Modeling as a Means of Integration Data on Substances Thermophysical Properties. Proc. of 15th All-Russian Science Conference "Electronic Libraries: Advanced Approaches and Technologies, Electronic Collections" – RCDL-2013, Yaroslavl, Russia, October 14–17, 2013. http://rcdl.ru/doc/2013/paper/s1_3.pdf

[2] Introduction to JSON. http://json.org/json-ru.html

[3] 3Vs (volume, variety and velocity), definition from TechTarget Network. http://whatis.techtarget.com/definition/3Vs

[4] Apache Hadoop. https://hadoop.apache.org/

[5] Apache Spark. http://spark.apache.org/docs/

[6] ChemSpider. www.chemspider.com

[7] Frenkel, M., Chirico, R.D., Diky V. et al.: XML-based IUPAC Standard for Experimental, Predicted, and Critically Evaluated Thermodynamic Property Data Storage and Capture (ThermoML) (IUPAC Recommendations 2006). Pure Appl. Chem., 78 (3), pp. 541-612 (2006)

[8] Hall, S.R, McMahon, B.: The Implementation and Evolution of STAR/CIF Ontologies: Interoperability and Preservation of Structured Data. Data Science J., 15 (3), pp. 1-15 (2016). doi: http://dx.doi.org/10.5334/dsj-2016-003

[9] Kaufman, J.G., Begley, E.F.: MatML. A Data Interchange Markup Language. Advanced Materials & Processes/November, pp. 35-36 (2003)

[10] Michel, K., Meredig, B.: Beyond Bulk Single Crystals: A Data Format for all Materials Structure-property-processing Relationships. MRS Bulletin. 41 (8), pp. 617-623 (2016)

[11] Ontobee: A Linked Data Server Designed for Ontologies. www.ontobee.org

Высокоуровневая формализация предметной области для консолидации информационных ресурсов в области неорганического материаловедения

© В.А. Дударев[1,2] © Н.Н. Киселева[1]
[1]Институт металлургии и материаловедения им. А.А. Байкова РАН,
[2]Национальный исследовательский университет «Высшая школа экономики»,
Москва, Россия

vic@imet.ac.ru kis@imet.ac.ru

Аннотация. Обоснована актуальность интеграции информационных систем по свойствам неорганических веществ и материалов. Отмечено, что консолидация возможна только на основе формализации предметной области. Введены основные определения и предложена формализация содержимого информационных систем по свойствам неорганических веществ и материалов на базе трех моделей: вербальной, теоретико-множественной и объектно-ориентированной.

Ключевые слова: интеграция информационных систем, неорганическая химия.

High-level Formalization of Problem Domain for Inorganic Materials Science Information Resources Consolidation

© V.A. Dudarev[1,2] © N.N. Kiselyova[1]
[1]Institution of Russian Academy of Sciences A.A. Baikov Institute of Metallurgy and Materials Science RAS,
[2]National Research University Higher School of Economics,
Moscow, Russia

vic@imet.ac.ru kis@imet.ac.ru

Abstract. Information systems on inorganic substances and materials properties integration actuality is grounded. It's noted that consolidation is possible on basis of subject domain formalization only. The paper introduces principal terms definitions and proposes high-level formalization of information systems on inorganic substances properties contents by means of three models: verbal, set-theoretical and object-oriented.

Keywords: information system integration, inorganic chemistry.

1 Введение

Современные исследования во многих областях науки отличаются интенсивным накоплением и обработкой больших массивов данных. Развитие неорганической химии, как науки, привело к огромному числу исследовательских работ, направленных на всестороннее исследование свойств различных классов неорганических веществ. Результаты этих исследований, как правило, оформляются в виде текстов научных работ, что на данном этапе развития информационных технологий (ИТ) делает практически невозможным компьютерный анализ и обработку имеющихся публикаций с целью извлечения из них знаний и фактов.

Труды XIX Международной конференции «Аналитика и управление данными в областях с интенсивным использованием данных» (DAMDID/ RCDL'2017), Москва, Россия, 10–13 октября 2017 года

Разработка специализированных информационных систем (ИС) по свойствам неорганических веществ и материалов (СНВМ) является необходимым для успешного развития многих наукоемких областей современной промышленности, например, электроники и машиностроения, т. к. позволяет выбрать оптимальные материалы для решения возникающих задач. Поэтому во многих развитых странах вкладываются значительные инвестиции в создание и развитие ИС СНВМ и расчетных систем, в том числе, на основе машинного обучения [1], которые являются по сути инфраструктурным фундаментом не только для инновационной промышленности, но и для самой науки о материалах.

2 Трудности доступа к информации по СНВМ

Необходимо отметить, что не существует ИС СНВМ, которая содержала бы все требуемые для

анализа данные, и часто информация распределена по нескольким ИС СНВМ, поэтому на практике доступ к такой распределенной по разнообразным источникам информации и ее всесторонний анализ даже для специалиста являются проблемой, решение которой неизбежно связано с двумя задачами. Во-первых, для поиска необходимой информации требуется знать, как минимум, перечень ИС СНВМ, в которых может содержаться искомая информация. Во-вторых, специалисту необходимо, имея доступ к целевым ИС, осуществить поиск необходимой информации и ее всесторонний анализ.

Решение первой задачи поиска нужной ИС СНВМ облегчается за счет использования специализированной ИС Information Resources on Inorganic Chemistry (IRIC), описывающей информационные ресурсы по неорганической химии и материаловедению. По своей сути IRIC является попыткой систематизации наиболее значимых ИС СНВМ [2]. Система реализована в виде веб-приложения и круглосуточно доступна по адресу http://iric.imet-db.ru/ на русском и английском языках.

Для решения второй задачи – обеспечения доступа к ИС СНВМ с возможностью быстрого поиска требуемой информации – необходима интеграция ИС в данной предметной области, что является не только большой организационной, но и технической проблемой.

3 Вербальное описание предметной области

Для успешной консолидации любых ИС необходимо, прежде всего, формализовать описание предметной области, которому должны соответствовать интегрируемые ИС.

Отличительной особенностью многих ИС СНВМ является узкая предметная направленность, обусловленная спецификой области исследования. Поэтому такие системы хранят информацию только о тех веществах и их характеристиках, которые относятся к исследуемой предметной области. В качестве примера можно привести ИС по фазовым диаграммам систем с полупроводниковыми фазами «Диаграмма» [3] и ИС по веществам с особыми акустооптическими, электрооптическими и нелинейнооптическими свойствами «Кристалл» [4]. Эти системы ориентированы на специалистов-материаловедов в области химии и материаловедения полупроводников и диэлектриков.

Таким образом, в разных информационных системах представлены различные характеристики (будем далее называть их свойствами) различных веществ и материалов (будем далее называть их сущностями). Значения свойств определяются, в первую очередь, составом неорганических веществ (набором химических элементов, входящим в их состав, и их соотношением, т. е. качественным и количественным составом), а также часто физические свойства зависят от кристаллической структуры образовавшейся твердой фазы. Поскольку ИС СНВМ тесно связаны с химией, то сущности в ИС СНВМ могут быть описаны с помощью иерархии понятий (система → вещество → модификация) в виде дерева (Рис. 1).

Обозначим сущности второго уровня общим термином «вещество», понимая под этим термином совокупность дискретных образований, обладающих массой покоя (т. е. атомы, молекулы и то, что из них построено). Итак, при описании химических сущностей можно использовать три уровня: система, вещество и кристаллическая (полиморфная) модификация (далее – модификация). При этом каждый последующий уровень *уточняет* (конкретизирует) информацию об описываемом химическом объекте.

Рисунок 1 Вершина иерархии понятий химических сущностей в неорганической химии

Приведем кратко определения основных терминов, использованных в иерархии понятий.

Химическая система (элементы, определяющие качественный состав) – система, образованная химическими элементами. Она может быть описана как множество атомов, образующих химическую систему. Более строго, химическая система – это совокупность микро- и макроколичеств веществ, способных под воздействием внешних факторов (условий) к превращениям с образованием новых химических соединений. Например, химическая система, в которую входят элементы медь, галлий и теллур, обозначается как Cu-Ga-Te.

Химическое соединение – однородное вещество постоянного или переменного состава с качественно отличным от свойств образующих его элементов химическим или кристаллохимическим строением. Соединение образовано из атомов нескольких химических элементов, связанных химической связью. На фазовой диаграмме область гомогенности соединения отделена (при всех температурах и давлениях) от области компонентов или твердых растворов на их основе. Элементы в соединении не могут быть разделены простым механическим способом, а лишь химической обработкой, нагреванием, электрическим током и т. д.

Раствор – макроскопически гомогенная смесь двух или более компонентов, состав которой при данных внешних условиях может непрерывно меняться в некоторых пределах.

Гетерогенная смесь – механическая смесь разнородных компонентов, в которой при заданных условиях отсутствует химическое взаимодействие.

Кристаллическая (полиморфная) модификация – форма пространственной организации твердого вещества.

Указанные выше химические определения являются в значительной степени нечеткими (размытыми). Поэтому иногда трудно провести

границу между, например, упорядоченным твердым раствором и соединением.

Необходимо отметить, что описание сущностей и их свойств в разных ИС по свойствам веществ происходит с разной степенью детализации. Так, например, в ИС «Диаграмма» описание большинства свойств химических сущностей ведется на уровне химических систем. А в ИС «Кристалл» некоторые свойства описаны на уровне химических веществ (например, температура плавления, растворимость и пр.), а некоторые свойства представлены на уровне конкретных модификаций (например, нелинейно-оптические коэффициенты, показатели преломления и пр.).

Очевидно, что свойства, указанные для химических сущностей на уровне систем, распространяются на все химические вещества этой системы и их модификации. Аналогично свойства, заданные на уровне химических веществ, распространяются на все химические модификации этого вещества. Данные замечания важны в контексте формального моделирования предметной области

4 Формальное описание предметной области в терминах теории множеств

При консолидации ИС возникают синтаксические и структурные конфликты из-за того, что ИС используют данные, различные по синтаксическому описанию и структуре. В ряде ИС используются реляционные системы управления базами данных (СУБД), в других – иерархические СУБД. В последнее время нередко строятся ИС, которые используют форматы JSON (JavaScript Object Notation), XML (eXtensible Markup Language) или какие-либо его известные приложения, например, RDF для хранения информации. В ИС, разработка которых велась довольно давно, нередко можно встретить собственные двоичные форматы для хранения и обработки данных. Все это многообразие моделей данных и схем представления, а также обработки информации приводит к тому, что ИС в том виде, в котором они существуют, зачастую являются несовместимыми с другими программными продуктами. Следует отметить, что изначально при проектировании ИС СНВМ взаимодействие с внешней программной средой не предусматривалось вовсе.

Разрешить синтаксические и структурные конфликты можно за счет введения общей схемы представления информации и обмена данными, построенной согласно описанию предметной области. Как уже было отмечено выше, при описании химических сущностей можно использовать три уровня: система, вещество и кристаллическая модификация. Указанная иерархия химических сущностей, которая рассматривается в контексте интегрированной ИС, представлена на Рис. 2.

Таким образом, в общую схему предметной области закладывается три типа объектов, соответствующих химическим сущностям: система (или химическая система – качественный состав вещества), вещество (количественный состав вещества) и модификация. При этом каждый последующий уровень *уточняет* (конкретизирует) описание объекта. Следовательно, все оболочки интегрируемых ИС СНВМ должны оперировать этими тремя типами объектов при ссылке на химические сущности. При этом стоит учитывать, что если характеризуется определенная кристаллическая модификация, то определена также и химическая система с веществом, модификация которого представляется, т.е. если описание химической сущности ведется на уровне модификаций, то все вышележащие уровни (вещество и система) считаются описанными. Следует заметить, что обратное неверно: при известном описании химической системы вещество и модификация не определены. Однако необходимо понимать, что при описании сущности на уровне системы все описанные свойства автоматически распространяются на все вещества и модификации, образованные в рамках этой системы. Это во многом напоминает наследование в объектно-ориентированном программировании (ООП).

Рисунок 2 Иерархия химических сущностей, рассматриваемая в контексте интегрированной ИС СНВМ

Воспользуемся теорией множеств для описания сущностей рассматриваемой предметной области, учитывая, что каждый последующий уровень в иерархии уточняет (дополняет) описание объекта. Обозначим множество химических систем S, множество химических веществ C, а множество кристаллических модификаций M. Тогда химическая система будет обозначаться s ($s \in S$), химическое вещество обозначим c ($c \in C$), а кристаллическую модификацию – m ($m \in M$).

Химическая система s может быть представлена как множество химических элементов e_i: $s=\{e_1,e_2,\ldots,e_n\}$. Химическое вещество c определяется не только множеством атомов (химических элементов), но и их количественным вхождением в состав вещества, раствора или смеси. Поэтому вещество c может быть представлено кортежем (s,f), где $s \in S$, а f является отображением множества атомов (химических элементов), которые образуют вещество, на множество пар $R^* \times R^*$, задающих соответственно минимальное и максимальное вхождения заданного элемента в вещество, раствор или смесь c. Значит, $f: e_i \rightarrow (R^*_{\min}, R^*_{\max})$, где $R^* = R^+ \cup \{x\}$.

R^+ – множество неотрицательных действительных чисел, а R^* – это множество R^+, расширенное элементом x. Элемент x служит для обозначения неизвестного числа, так как при обозначении смесей, где вхождение компонентов может варьироваться, принято использовать x для обозначения неизвестного, например, $Fe_{1-x}Se_x$. R^*_{min} и R^*_{max} – соответственно, минимальная и максимальная концентрации химического элемента e_i в веществе c. В случае, когда концентрация конкретного химического элемента e_i в веществе c фиксирована, $R^*_{min}=R^*_{max}$. Химическая модификация m может быть представлена кортежем (s, f, mod), где $s \in S$, $f: e_i \rightarrow (R^*_{min}, R^*_{max})$, а mod – строковое обозначение кристаллической модификации вещества, принятое в интегрированной ИС (одно из значений перечисления (enum) сингоний: {*Triclinic, Monoclinic, Orthorhombic, Tetragonal, Trigonal, Hexagonal, Cubic*}).

5 Формальное описание предметной области на объектно-ориентированном языке

При использовании объектно-ориентированного языка достаточно просто могут быть описаны формализмы предметной области, описанной выше. В качестве подтверждения данного тезиса рассмотрим формализацию с использованием языка C# (свободно доступная версия 6.0 https://github.com/vicdudarev/ChemicalHierarchy).

Не рассматривая детально предложенную реализацию, остановимся кратко на переходе от системы к веществу – дополнении информации о качественном составе количественным описанием. В предлагаемой реализации химическая система (класс ChemicalSystem) описывается в качестве одномерного массива типа ChemicalElement[], где ChemicalElement – класс для представления химического элемента (содержит обозначение элемента и его атомный номер). На уровне описания количественного состава вводится наследуемый от ChemicalSystem класс ChemicalSubstance, расширяющий описание количественным составом, представленным в виде одномерного массива типа Quantity[], где Quantity – простейший класс, содержащий пару значений Min и Max. Отметим, что в конструкторах классов выполняются все проверки на корректность задаваемых значений. Например, в конструкторе объектов класса ChemicalSubstance проверяется, что размер массива количественного описания совпадает с размером массива качественного описания, унаследованного от ChemicalSystem. Таким образом, развитые возможности объектно-ориентированных языков позволяют корректно реализовать предлагаемую в разделе 4 формализацию.

6 Представление свойств сущностей

Рассмотрев формализацию описания химических сущностей, перейдем к краткому изложению предлагаемого представления свойств химических сущностей. Как было отмечено, в интегрируемых ИС содержится информация по свойствам химических сущностей, например, плотность, растворимость, теплопроводность, ширина запрещенной зоны и т. п. При этом для каждой химической сущности в базе данных (БД) ИС нередко содержится несколько записей для описания значения свойства. Это обусловлено разными обстоятельствами. Во-первых, информация, содержащаяся в БД ИС, может быть взята из различных источников, при этом данные нередко расходятся. Это объясняется различными способами измерения, точностью измеряющей аппаратуры и т. д. Таким образом, в ИС СНВМ приводится несколько вариантов значения, например, плотности соединений. Во-вторых, значения рассматриваемых свойств зачастую зависят от внешних условий, при которых проводились измерения. Например, такие параметры, как растворимость и ширина запрещенной зоны, зависят от температуры. Другими словами, свойства часто являются функциями от различных аргументов, число которых, строго говоря, не фиксировано. Это означает, что разные свойства могут иметь разную структуру представления данных. Более того, одно и то же свойство в разных ИС СНВМ может фактически являться функцией от разного числа аргументов, и поэтому невозможно будет предложить универсальный формат представления заданного свойства для всех ИС. Это во многом может быть объяснено тем фактом, что при детальном исследовании какого-либо свойства число таких функциональных зависимостей от внешних параметров может возрастать. Следовательно, если такое свойство будет подробно рассмотрено в некоторой ИС СНВМ, которая еще не включена в общую интегрированную ИС, то при ее включении в состав интегрированной ИС возникнет проблема согласования форматов представления указанного свойства. Таким образом, невозможно заранее предусмотреть все зависимости и заложить их в общий формат представления данных для даже отдельно взятого конкретного свойства, не говоря о представлении свойств в целом.

В связи с вышеуказанным необходим некоторый механизм, позволяющий гибко представлять значения свойств в рамках интегрированной ИС. В настоящее время существует ряд широко используемых языков описания произвольных форматов данных, среди наиболее распространенных – JSON и XML. С помощью этих языков удобно описывать различные структуры данных, они являются межплатформенными форматами и поддерживаются большинством языков и библиотек [5]. На сегодняшний день представление данных с помощью таких языков является фундаментом для обеспечения взаимодействия различных программно-аппаратных платформ. В настоящее время все большее количество информации в современных промышленных системах представляется в форматах JSON и XML.

Использование этих форматов является целесообразным еще и потому, что они используются в качестве основы функционирования веб-сервисов.

Для разрешения семантических и структурных конфликтов необходимо стандартизировать форматы представления описанных химических сущностей и свойств в рамках интегрированной ИС на языках XML и JSON, т. е. необходимо разработать форматы соответствующих документов для представления химических сущностей, их свойств и другой информации. Это позволит обмениваться информацией между звеньями интегрированной ИС.

7 Заключение

Проблема интеграции ИС вообще и ИС СНВМ, в частности, чрезвычайно актуальна, поскольку доступ ко всей совокупности данных о веществах позволяет рассматривать такой консолидированный информационный источник в качестве объекта для всестороннего анализа и извлечения новых знаний.

В неорганическом материаловедении на первом этапе наиболее реалистичными являются попытки интеграции, основанные на учете специфики предметной области. Предложенное выше формальное описание предметной области – неорганического материаловедения – ни в коем случае не претендует на глубину проработки, которая бы удовлетворила материаловеда. В каждой из многочисленных областей материаловедения существует множество своих особенностей, учесть которые в большей или меньшей степени возможно при построении онтологий этих областей, основанных на сложных таксономиях.

Важно понимать, что сложность реализации ИС напрямую зависит от сложности формального описания предметной области. В этом смысле предложенная формальная модель (система → вещество → модификация), на наш взгляд, является приемлемым компромиссом между сложностью реализации интегрированной ИС и детальностью описания информации, представленной в отдельных интегрируемых ИС СНВМ.

Поддержка

Работа выполнена при частичной финансовой поддержке РФФИ, проекты 16-07-01028, 17-07-01362 и 15-07-00980.

Литература

[1] Киселева, Н.Н.: Компьютерное конструирование неорганических соединений. Использование баз данных и методов искусственного интеллекта. М.: Наука (2005)

[2] Киселева, Н.Н., Дударев, В.А.: Информационная система по ресурсам неорганической химии и материаловедения. Вестник Казанского технологического университета, 17 (19), сс. 356-358 (2014)

[3] Христофоров, Ю.И., Хорбенко, В.В., Киселева, Н.Н. и др.: База данных по фазовым диаграммам полупроводниковых систем с доступом из Интернет. Изв. вузов. Материалы электронной техники, (4), сс. 50-55 (2001)

[4] Киселева, Н.Н., Прокошев, И.В., Дударев, В.А. и др.: Система баз данных по материалам для электроники в сети Интернет. Неорган. материалы, 42 (3), сс. 380-384 (2004)

[5] Christophides, I., Koffina, G., Serfiotis, V, Tannen, A.: Integrating XML Data Sources using RDF/S Schemas: The ICS-FORTH Semantic Web Integration Middleware (SWIM), Deutsch Dagstuhl Seminar: Semantic Interoperability and Integration (2004)

Integrating Data Analysis Tools for Better Treatment of Diabetic Patients

© Svetla Boytcheva[1] © Galia Angelova[1] © Zhivko Angelov[2] © Dimitar Tcharaktchiev[3]

[1] Institute of Information and Communication Technologies, Bulgarian Academy of Sciences, Sofia, Bulgaria
[2] Adiss Lab Ltd., Sofia, Bulgaria
[3] Medical University Sofia, University Specialized Hospital for Active Treatment of Endocrinology, Sofia, Bulgaria

svetla.boytcheva@gmail.com, galia@lml.bas.bg, angelov@adiss-bg.com, dimitardt@gmail.com

Abstract. This paper presents the construction and usage of an anonymous Diabetes Register for patients in Bulgaria. The Register is generated automatically from outpatient records submitted to the Bulgarian National Health Insurance Fund in 2010–2014 and continuously updated using outpatient records for 2015-2016. The construction relies on advanced automatic analysis of free text information as well as on Business Analytics technologies for storing, maintaining, searching, querying and analyzing data. Original frequent pattern mining algorithms enable to find patterns and sequences taking into account temporal information. The paper discussed the software environment as well as experiments in frequent pattern mining that enable knowledge discovery in the very large repository underlying the Register (currently 262 million pseudonymizedoutpatient records submitted to the Bulgarian National Health Insurance Fund in 2010–2016 for more than 5 mln citizens yearly). The claim is that the synergy of modern analytics tools transforms a static archive of clinical patient records to a sophisticated software environment for knowledge discovery and prediction.

Keywords: synergy of data management, data mining and text mining tools; clinical data; frequent pattern mining; data analytics; natural language processing; knowledge discovery.

1 Introduction

Medicine is known as a Data Intensive Domain: due to the recent penetration of the Information and Communication Technologies (ICT) in all areas of our society, a rapidly increasing amount of medical data is produced by the healthcare sector, on the one hand, and by biomedical research on the other hand. In the healthcare sector, ICT applications support health diagnostics, development and maintenance of medical Electronic Health Records, telemedicine and telecare, patient administration, almost all aspects of healthcare management and healthcare delivery as well as medical education and training. In biomedical research, progress in deeper understanding of medical phenomena is sought by construction of big data models: e.g. virtual physiological human, models of brain, in computational genetics and so on. Public access to health information is changing the relationship between the patients and the health institutions that are responsible for care delivery. The monitoring and control function of patient organizations is facilitated by the modern ICT tools as well. Today we are still in an early phase of a long-term technological and social shift that will be implied by advancing further the ICT fundamentals and tools.

In this paper we present the integration of various ICT tools for automatic generation of a Diabetes Register for Bulgarian patients. The huge amount of clinical data, underpinning a repository of Outpatient Records (ORs), enabled to construct interfaces that support both *monitoring* functionalities (oriented to the health management authorities) and *research*-oriented functionalities for knowledge discovery. The monitoring functionalities are based on business analytics while research tools use data mining and pattern search. The software environment includes also components for automatic analysis of free texts in Bulgarian. These components facilitate the Register generation and its update because they deliver values of clinical tests and lab data which are described as unstructured text only.

This paper is structured as follows. Section 2 overview related work in several areas that are relevant to the subject: Diabetes registers, Natural Language Processing (NLP) for clinical narratives, Business Intelligence (BI) and analytics, Frequent Pattern Mining (FPM). Section 3 presents the experimental study context and summarizes the developments during the last 3-4 years (because the Register was built iteratively). Section 4 presents relevant achievements in automatic analysis of clinical narratives in Bulgarian language. Section 5

Proceedings of the XIX International Conference "Data Analytics and Management in Data Intensive Domains" (DAMDID/RCDL'2017), Moscow, Russia, October 10–13, 2017

discusses recent algorithms for frequent pattern mining and presents experiments related to knowledge discovery in the Register repository. Section 6 contains the conclusion and plans for future work.

2 Related Work

2.1 Diabetic Registers

There are several nation-wide Diabetes Registers in the world, e.g. in Denmark [1], Sweden [2], Norway [3]. Registers explicate the number of patients who are diagnosed with Diabetes and provide good monitoring and control. Constructing registers is expensive and burdening the patients as well as the medical experts with additional administrative work. Furthermore, in some countries chronic disease management is not recognized as a part of general medical practice. As for the construction, most medical experts agree that Registers are a must since Diabetes is a chronic disease with significant social consequences. Electronic patient registration systems are proposed like the one in Ireland [4] (but it is not implemented yet). It is interesting to mention that in Sweden, during the Diabetic Register development phase 2001–2005, the registration rate of patients gradually increased and reached 75% which in 2010 still remains stable and is one of the highest in the country [5].Thus infrastructure construction is a critical issue but data collection and update are further problems that can be solved only by persistency and diligence.

2.2 NLP ofClinical Narratives

Usually automatic analysis of clinical narratives is implemented partially: only fragments of the text are considered. The phrases, selected as "interesting", are typically picked up due to the presence of a word or an entity which are considered "significant". This approach for shallow analysis is called "Information Extraction" (IE). IE from clinical texts matures only recently but its accuracy gradually improves and often exceeds 90% [6]. The review [6] stated in 2008 that "current applications are rarely applied outside of the laboratories they have been developed in, mostly because of scalability and generalizability issues". Today, however, this is valid for languages other than English because, with the active contribution of numerous research groups in the USA, NLP for English clinical narratives has much better performance at present. Comprehensive language resources exist for English, such as UMLS [7] as well as tools like Knowledge Map Concept Identifier [8] which processes clinical notes and returns CUIs (Concept Unique Identifiers) for the recognized UMLS terms. Another important tool is the public NegExsystem which identifies and interprets negations in English clinical texts [9,10]. We also mention the open-source cTAKES[19] (clinical Text Analysis and Knowledge Extraction System) and the Health Information Text Extraction (HITEx) system [11]. A recent study [12] enumerates the advantages to incorporate NLP for English in medical systems: it systematically links several terms to a concept using databases that standardize health terminologies; avoids manual work for searching term variations; increases the number of patients in the considered cohorts and thus increases the sensitivity of the recognition. Despite the NLP limitations, the conclusion is that NLP engines are powerful components ready for integration in medical data mining and – due to improvements expected in the future, e.g. more accurate mappings of terms to medical concepts – the importance of NLP as a valuable supporting technology will grow.

Here we consider NLP for Bulgarian clinical text. No comprehensive resources exist for Bulgarian medical language; the International Classification of Diseases ICD-10 is the only terminological resource which is available in electronic format. Our experience shows that within 2-3 years one can achieve good performance in separate extraction tasks. We apply software prototypes developed some years ago that are gradually improved. The most useful tools area drug extractor (it finds in the free text the drug name, dosage, frequency and route of admission [13]) as well as an extractor of numeric values of lab data and clinical tests [14].

2.3 Big Data, Business Intelligence Tools

Big Data usually designates a massive volume of structured and unstructured data, too large or too dynamic to be processed by traditional software tools and techniques. The popular "3Vs" features of Big Data were first introduced by Gartner (previously META group): "high Volume, high Velocity, and/or high Variety" [15]. Wikipedia is an example for big data consisting of unstructured texts, images and hyperlinks. Big data analytics is the process of collecting, organizing and analyzing big data to discover useful information. Business Intelligence tools analyze big data of enterprises in order to provide historic, present and predicted views to the business processes. Predictive analytics for establishment of trends is the preferred functionality in contrast to databases that deal with data items and extract subsets of data values. Visualization is an important feature of BI tools because they show generalizations and tendencies in one screen [16]. Another necessary feature is the speed of processing since big data often appear in real time.

In our project we use a BITool which stores data in *n*-dimensional cubes and explores multi-dimensional data i.e. hyper planes [17]. The user can split the dataset into groups of objects with similar features. If temporal dimension is included the user can track changes of object characteristics over time by animation. BITool enables the discovery of similar situations over time when a search pattern is specified for a particular period.

2.4 Frequent Pattern Mining

There are two principal tasks in pattern search: frequent pattern mining (FPM) where the events (objects) are considered as unordered sets, and frequent sequence mining (FSM). Approaches for solving the

[19]http://ctakes.apache.org/

FPM task vary from the naïve Brute Force and Apriori algorithms, where the search space is organized as a prefix tree, to Eclat algorithm that uses tid sets directly for support computation by processing prefix equivalence classes [18]. Most FPM and FSM methods do not consider contextual information about extracted patterns. They usually build a (huge) prefix tree. Most FPM algorithms generate all possible frequent patterns (FPs). Summarized information for data relations can be extracted as maximal frequent item sets (MFI) in order to reduce redundancy and decrease significantly the number of FPs for post-analysis. All classic algorithms for FPM can be modified for MFI search.

We have proposed a novel algorithm for mining sets of events in order to identify strong co-occurrence of patterns [19]. It is a cascade data mining approach for FPM enriched with context information which aims at the discovery of complex relations between medical events with respective timestamps. Experiments with this approach are presented in Section 5 to illustrate the functionality of the Diabetes Register as a research tool.

3 Experimental Study

3.1 Principal Objective

A pseudonymized Register of diabetic patients was generated in 2015 from the Outpatient Records, collected by the Bulgarian National Health Insurance Fund (NHIF), in compliance with all legal requirements for safety and data protection [20]. The usual patient registration process was kept without burdening the medical experts with additional paper work. NHIF is the only obligatory Insurance Fund in Bulgaria so we note that working with ORs ensures 100% registration of all patients who contacted the healthcare system at all (however there are Bulgarian citizens who are not insured and some others who have ORs but are not properly diagnosed with Diabetes). The data repository, underpinning the Register, currently contains more than 262 mln pseudonymised ORs submitted to the NHIF in 2010-2016 for more than 7.3 mln Bulgarian citizens (more than 5 mln yearly), including 483,836 diabetic patients. In Bulgaria ORs are produced by General Practitioners (GPs) and Specialists from Ambulatory Care whenever they contact patients. Despite the primary accounting purpose ORs summarize sufficiently the case and motivate the requested reimbursement. They are semi-structured files with predefined XML-format. Many indicators in the Register copy the structured data submitted to NHIF in ORs: *(i)*date and time of the visit;*(ii)* pseudonymized personal data, age, gender;*(iii)* pseudonymised visit-related information; *(iv)* diagnoses in ICD-10; *(v)* NHIF drug codes for medications that are reimbursed; *(vi)* a code if the patient needs special monitoring; *(vii)* a code concerning the need for hospitalization; *(viii)* several codes for planned consultations, lab tests and medical imaging.

ORs contain also important values presented in free text fields: glycated haemoglobin (HbA1c), body mass index (BMI), weight, blood glucose and blood pressure etc. These values are essential for a Diabetic Register so they are extracted automatically from four XML fields: *(i) Anamnesis:* summarizes case history, previous treatments, often family history, risk factors; *(ii)Status*: summary of patient state, height, weight, BMI, blood pressure etc.; *(iii)Clinical tests*: values of clinical examinations and lab data listed in arbitrary order; *(iv)Prescribed treatment:* codes of drugs reimbursed by NHIF, free text descriptions of other drugs. Integration of large scale text analysis is a real novelty in this field.

3.2 Analytics Using BITool

Today the system BITool supports the Diabetes Register at the University Specialized Hospital for Active Treatment of Endocrinology "Acad. Ivan Penchev", Medical University – Sofia (this Hospital was authorized by the Bulgarian Ministry of Health to host the Register of diabetic patients in Bulgaria). BITool's functionalities enable the monitoring of significant indicators like glycated hemoglobin (HbA1c) and blood glucose values. In this way the Register achieves its objective: to provide an adequate monitoring strategy for diabetic patients and to improve the healthcare and quality of life for the patients and their families. Two examples illustrate the services. Figure 1 shows the number of diabetic patients in the dimensions age-gender (at certain moment). Here BITool operates on the structured information from the NHIF archive: patient pseudonym, age and gender. Further statistics of this kind might concern explorations of diabetic patients per region code, types of diabetes and diabetes complications, per GPs, per types of medication, according to frequency of visits etc.

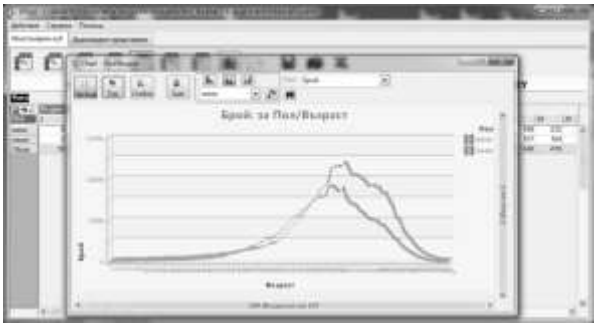


Figure 1 Number of diabetic patients grouped by age

Figure 2 explores the tendency in the development of treatment. It displays the number of patients who had changes in the HbA1c levels within the interval [–5,5] units for certain period of time. For most patients the HbA1c level decreased by 1 unit. The HbA1c levels are extracted from the free text of ORs for the corresponding patients with timestamp.

Finally we show the Register interface during the process of exploring the collection of ORs (Figure 3). The names and personal identifiers of patients and GPs are replaced by pseudonyms; only the name of the city/village remains in the address field.

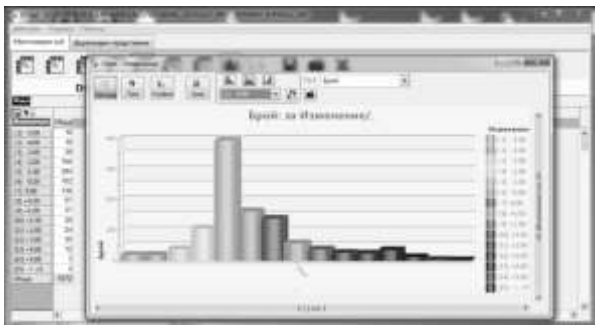


Figure 2 Reduction of HbA1c levels after application of incretin[20] based drugs

Figure 3 Exploring outpatient records in the Register

4 NLP for Bulgarian Clinical Narratives

Design and implementation of software for automatic extraction of patient-related entities from a Big Data collection is a quite challenging task. One needs to scale up existing research prototypes to process millions of patient records, coping with noisy and missing data, and still providing reliable results. Some numeric entities refer to key risk factors for development of Diabetes Mellitus (levels of glycated hemoglobin HbA1c and blood glucose) and cardio-vascular diseases (high blood pressure). Unfortunately in the Bulgarian clinical practice these values are usually documented in free text paragraphs, presented in a huge variety of formats, so their automatic identification is difficult. We note that according to some studies, today more than 80% of the patient-related clinical information is stored as free text in the Electronic Health Record systems.

In [21] we proposed a hybrid method for automatic generation of grammar rules for IE from clinical data. The experiments were made and evaluated over approximately 9.5 million of ORs. Here we cite only the evaluation of blood pressure extraction from the ORs of about 1,800,000 patients with arterial hypertension for 3 year period: all available values are about 38.3 million and the extraction was performed with precision 92% and recall 98%. The variety of recording formats and explanations written by thousands of medical professionals require constant evaluation of grammar coverage and extraction accuracy in general. Some of the main advantages of the proposed method, beyond its reliable performance and good precision in text mining, are the modularity, extensibility, and scalability.

[20]https://www.drugs.com/drug-class/incretin-mimetics.html

5 Researches in Frequent Pattern Mining

5.1 Contextual Information

Most FSM and FPM approaches do not use contextual information about extracted patterns. These algorithms extract general templates but do not answer the major question whether they are influenced in some way by the context and whether they are valid in various aspects. Existing methods which search for patterns using contextual information are based on attributes that are organized into hierarchical structures and on attributes' generalizations and specializations.

Context information is organized as attributes of item sets and tid sets. Attributes may have different organization – structured or unstructured. This enables to explore the context-dependent templates. Rabatel et al.[22] propose an approach in marketing domain taking into account not only the transactions that have been made but also various attributes associated with customers like age, gender etc. Attributes have a hierarchical structure ($H(Age), H(Gender)$) and explore patterns at different levels of attributes abstraction–lattice H (Figure 4). Traditional methods consider only the top level $[*,*]$ – for any age and regardless of gender, i. e. without attributes. Rabatel et al. designed the algorithm Gespan and made experiments with about 100,000 product descriptions from *amazon.com.*

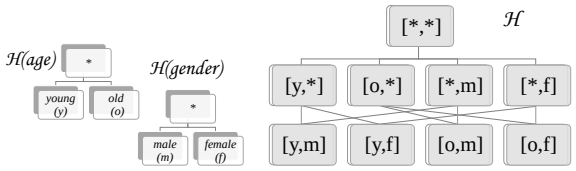


Figure 4 Structuring attributes in marketing domain

Ziembiński [23] proposes a new approach for extracting small contextual models from smaller collections of data that later are summarized in generalized models using information from contextual models with common information. This approach applies a metrics for measuring distance of context models. All values for similarity assessment are normalized in the range between 0.0 and 1.0. Attribute values are considered identical if the similarity function returns 1.0. In the opposite case the result is 0.0. This approach allows extracting patterns for data that would otherwise have to be dropped out of the templates because of its dispersion and low frequency.

5.2 Experimental Setup

We apply a retrospective analysis for patients from the Diabetes Register with Diabetes Type 2. The period of interest is two years preceeding the onset of the Diabetes Type 2, i. e. the so called prediabetes condition. In order to illustrate the potential of contextualized FPM we present results in searching comorbidities for patients in prediabet condition. Text mining modules are used to convert raw text descriptions to structured event data.

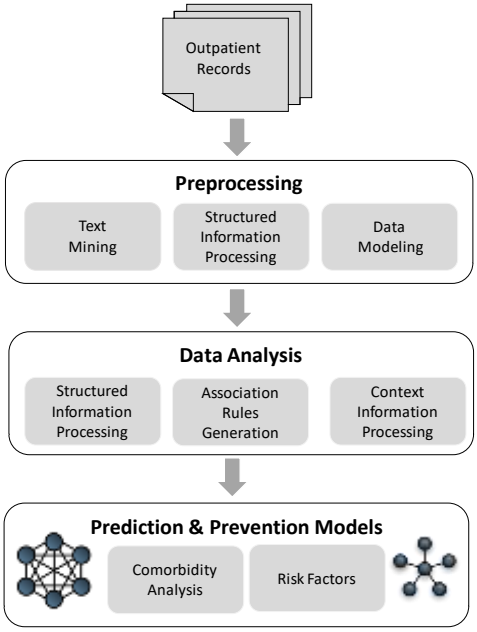


Figure 5 System Architecture

The search space is very large: the database is big, the number of diseases is also large. We propose a tabular method using a vertical database, depth-first traversal as well as set intersection and diffsets [19]. Further processing of the maximal frequent item sets (MFI) is applied to remove diagnostic-related groups. In addition some context information is added toeach MFI toinvestigate comorbidities. Furthermore association rules with lift are generated. The context information is represented as attribute-value tuples for each patient; the post-processing identifies the importance of different attributes for each MFI.

The architecture of the experimental workbench is shown on Figure 5. Our research [19] aims to develop further the ideas of the two contextual approaches for data mining [22, 23].

For the collection *S* of ORs we extract the set of all different patient identifiers $P = \{p_1, p_2, \cdots, p_N\}$. This set corresponds to transaction identifiers (*tids*) and we call them *pids* (patient identifiers). We consider each patient visit to a doctor as a single event. For each patient $p_i \in P$ an event sequence of tuples $\langle event, timestamp \rangle$ is generated: $E(p_i) = \big((e_1, t_1), (e_2, t_2), \ldots, (e_{k_i}, t_{k_i})\big), \quad i = \overline{1, N}$. Let $\mathcal{E}$ be the set of all possible events and $\mathcal{T}$ be the set of all possible timestamps. Let $I = \{id_1, id_2, \ldots, id_p\}$ be the set of all diseases ICD-10[21] codes, which we call *items*. Each subset $X \subseteq I$ is called an *itemset*. We define a projection function $\pi: (\mathcal{E} \times \mathcal{T})^N \to 2^I$: $\pi(E(p_i)) = I(p_i) = \big(id_{1i}, id_{2i}, \ldots, id_{m_i}\big)$, such that for each patient $p_i \in P$ the projected time sequence contains only the first occurrence (onset) of each disorder recorded in $E(p_i)$. Let $D \subseteq P \times 2^I$ be the set of all itemsets in our collection after projection π in the format $\langle pid, itemset \rangle$. We shall call D a *database*. We are looking for itemsets $X \subseteq I$ with frequency ($\sup(X)$) above given $minsup$. Let $\mathcal{F}$ denote the set of all frequent itemsets, i. e.

$$\mathcal{F} = \{X \mid X \subseteq I \text{ and } \sup(X) \geq minsup\}.$$

A frequent itemset $X \in \mathcal{F}$ is called *maximal* if it has no frequent supersets. Let $\mathcal{M}$ denote the set of all maximal frequent itemsets, i. e.

$$\mathcal{M} = \{X \mid X \in \mathcal{F} \text{ and } \nexists\, Y \in \mathcal{F}, \text{such that } X \subset Y\}.$$

Let denote the power set (set of all subsets) of itemset X. Then each subset of $X \in \mathcal{F}$ is also a frequent itemset, i. e. $\forall\, Y \in 2^X \text{ implies that } Y \in \mathcal{F}$. For each item $id \in I$ we define the set called *pidset*:

$$p(\mathbf{id}) = \{p_i \mid \langle p_i, I(p_i) \rangle \in D \text{ and } id \in I(p_i)\}.$$

To study the nature of comorbidities we need to investigate the context in which they occur. Therefore we add some semantic attributes to each event – demographics of patients, age and gender, treatment, status, lab data and etc.

We define a set of attributes of interest $A = \{a_1, a_2, \ldots, a_k\}$. Context *Q* for some patient $p_i \in P$ is defined as the set of attribute-value pairs from patient profile information: $Q(p_i) = \{(a_1, q_1), (a_2, q_2), \ldots, (a_k, q_k)\}$. In order to decrease the number of possible values of attributes we apply some aggregation of data. For instance age value is categorized according to the World Health Organization (WHO) standard age groups. Data for body mass index (BMI) are also categorized according to the WHO[22] standard classification – *underweight*, *normal weight*, *overweight*, *obesity*.

For some data concerning demographic information, like region ID we have large number of distinct values. For such data we add also some additional properties concerning background information for the region – e.g. whether it is *south*, *north*, *west*, *east*, *central*, *northwest* etc., and *mountain*, *river*, *sea*, *thermal spring*, *urban region* etc. For status and clinical test data we take the worst value for the period, according to the risk factors deinition.

In primary interest for Diabetes Type 2 are BMI, glycated haemoglobin, blood pressure (RR – Riva Roci), blood glucose, HDL-cholesterol.

From $Q(p_i)$ we generate a feature vector $v(p_i) = (v_{1i}, v_{2i}, \ldots, v_{mi})$, where each attribute $a_j \in A$ with N_j possible values is represented by N_j consecutive positions in the vector. For the set of maximal frequent itemset $\mathcal{M}$ with cardinality $|\mathcal{M}| = K$ we have classes of comorbidities. We apply classification of multiple classes in order to generate rules for each comorbidity class. We use large scale multi class classification because we deal with a big database and a large group of comorbidity classes. We use Support Vector Machines (SVM) and optimization based on block minimization method described by Yu et al. [24].

[21] International Classification of Diseases and Related Health Problems 10th Revision. http://apps.who.int/classifications/icd10/browse/2015/en

[22] WHO, BMI Classification http://apps.who.int/bmi/index.jsp?introPage=intro_3.html

Table 1 Data analysis results for patients in prediabetes condition

Set	2013		2014		2013-2014	
items	ICD-10 3 signs	ICD-10 4 signs	ICD-10 3 signs	ICD-10 4 signs	ICD-10 3 signs	ICD-10 4 signs
Patients	27,082	27,082	27,902	27,902	29,205	29,205
Outpatient Records	267,194	267,194	296,129	296,129	556,323	556,323
ICD-10 codes	1,142	4,701	1,145	4,834	1,257	5,503
minsup	0.01	0.01	0.01	0.01	0.01	0.01
Total MFI	203	486	219	512	521	1,406
Longest MFI	5	8	5	9	6	9
Frequent Itemsets	608	7,452	689	8,935	1,909	32,093
Association Rules	686	58,299	810	78,052	2,722	381,012

5.3 Experiments and Results

We report results for patients with Diabetes Type 2 onset in 2015.The ORs of these patients for the period 2013-2014 were excerpted from the Diabetes Register when, as we assume, these patients were in a pre-diabetes condition. The idea of this experiment is to check whether we can successfully discover risk factors for these patients looking only at their ORs in 2013 and 2014. Then, maping our hypotheses to the real data for 2015, we test whether our approach is reasonable. (We note that due to the relatively short period of observation and lack of data about mortality, at the moment we cannot follow diabetes development in longer periods.)

In the Register each OR, corresponding to a single visit, cointains up to 4 diagnoses encoded in ICD-10. Some diagnoses are presented by 4-sign encodings, i.e. in a more specific way, while others use the more general 3-sign encoding. Due to the hierarchical organization of ICD-10 we shall analyse individually two collections: the original one, that is more specific (with 4-sign codes - see Example 1) and we shall generalise also all diagnoses to more general classes (with 3-sign codes - see Example 2). The examples present collections of diagnoses for a patient with ID 2196365.

Example 1:
$I(2196365)$={I10, M10.9, M10, K76.9, K76, L94.1, L94, M06.9, G57.9, Z00.8, H53, M51.1, M33.9}

Example 2:
$I(2196365)$={I10, M10, K76, L94, M06, M51, M33, H53, Z00, G57}

For some patients, the available ORs contain no information about certain attributes of the context information (Table 2). It is well known that missing data in medical documentationis inevitable. Thus some attribute values are replaced by the value NA, which is considered as the most general value.

For example the context information for the patient with ID is: $Q(2196365) = \{(age, 58), (gender, 1), (region, 03), (bmi, 29.32), (hba1c, NA), (blood\_glucose, 6.39), (hdl\_cholesterol, 1.15)\}$.

Table 2 Data for attributes in the collections

A	attribute	2013	2014	2013-2014
a_1	*age*	27,082	27,902	29,205
a_2	*gender*	27,082	27,902	29,205
a_3	*region*	27,082	27,902	29,205
a_4	*bmi*	21,659	22,413	27,928
a_5	*HbA1c*	153	238	370
a_6	*HDL cholesterol*	4,917	4,815	6,952
a_7	*bloodglucose*	11,925	12,185	17,016

One of the generated Maximal Frequent Item sets (comorbidity class), whose support contains the pid=is:

MFI#12: Z00 I10 M51 #SUP: 453

The following charts show the distribution of patients in the support of "MFI#12" according to their age (Figure 6), gender (Figure 7), BMI (Figure 8), and the HDL Cholesterol (Figure 9) correspondingly.

We can observe that most patients in this support set have higher risk of Diabetes Type 2, due to the presence of multiple risk factors as obesity, medium or high levels of cholesterol and hypertension (diagnose with ICD-10 code I10).

Data about HbA1c are available only for 3 out of 453 patients, that is why we consider this attribute as a more general value ANY. But we note that the lack of HbA1c measurements is not surprising because tests for HbA1c are made when the Diabetes is diagnosed (and this has happened in 2015 for the selected patient cohort).

Data for blood glucose are available only for 30% of these patient and for 50% of them the values were high.

Deeper analyses reveal medical arguments why higher risk exist especially for the patients in the support set of MFI#12: Z00 I10 M51 #SUP: 453. The diagnose with ICD-10 code M51 (Thoracic, thoracolumbar, and lumbosacral intervertebral disc disorders) means that the patients have lower motor activity and sedentary lifestyle, which causes obesity, overweight, higher values of cholesterol and blood pressure and therefore increases the risk of developing Diabetes. Actually this has happened in 2015.

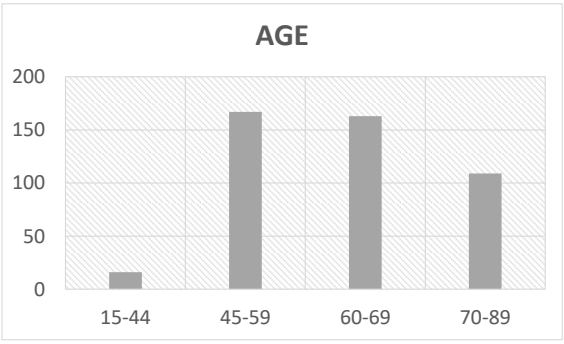


Figure 6 Age of the patients in the support set of "MFI#12"

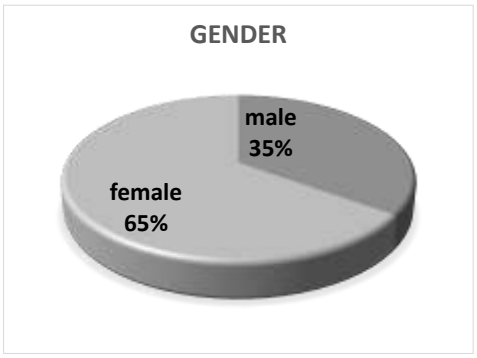


Figure 7 Gender of the patients in the support set of "MFI#12"

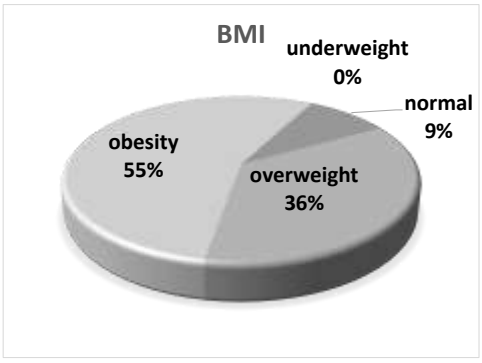


Figure 8 BMI of the patients in the support set of "MFI#12"

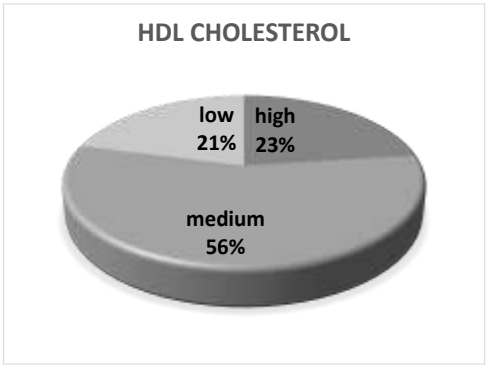


Figure 9 Levels of HDL Cholesterol of the patients in the support set of "MFI#12"

We note that in general the ICD-10 diagnose M51 is not considered risky for Diabetes. But our algorithm reveals this unknown and latent interrelationship.

6 Conclusion and Future Work

In this paper we present a software environment for collection and processing of Big Data in medicine - a Data Intensive Domain. The Diabetes Register has been developed stepwise and its research functionality is still under construction. We believe that the integration of various technologies is the proper way to approach the challenges of large-scale information processing because the integration ensures flexible multi-functionality and enables reuse of results.

The nation-wide Diabetes Register of Bulgaria is now visible in Internet23together with some public statistical information. We plan to develop the Register further as a predictive and preventing tool using the synergy of advanced technologies which enable to discover risk groups of patients that have predisposition to various socially-significant diseases. We have shown here that the present software environment is mature enough to identify patients with complexes of risk factors for development of Diabetes, e.g. risks like: family history (relatives with Diabetes); obesity; arterial hypertonia (RR>140/90); low physical activity; giving birth to a baby with weight more than 4 kg or gestational Diabetes; established impaired fasting glycaemia or impaired glucose tolerance; other states of insulin resistance (e.g. acanthosis nigricans, a specific hyperpigmentation of the skin that might be due to endocrine disorders); HDL-cholesterol≤0.90 mmol/l ortriglycerides≥2.2 mmol/l (≥2.82 mmol/l according to ADA);diagnosed polycystic ovarian syndrome, a cardio-vascular disease, or mental disorder setc. These risk factors are explicated in the patient-related documents either by values of clinical tests or by keywords and typical phrases that describe the factor. The patients with predisposition suffer from disorders and syndromes, diagnosed by various medical specialists in various time periods, but without any chance to establish connections between the medical doctors – e.g. a connection between a Psychiatrist and a Cardiologist that have consulted the patient. Elaborating further the analytics facility of the Register will provide functionality to monitor patient status over time, in the context of all available information, and to issue alerts for coincidence of risk factors that open the door to Diabetes and other chronic diseases. In this way we believe that in the foreseeable future it will become possible to identify the Bulgarian citizens who have predisposition to develop Diabetes Mellitus.

Acknowledgements

The research presented here is partially supported by the grant 02/4*SpecialIZed Data MIningMethoDs Based on Semantic Attributes* (IZIDA), funded by the National Science Fund in 2017–2019. The support of Medical University – Sofia, the Bulgarian Ministry of Health and the National Health Insurance Fund is acknowledged.

[23]http://usbale.com/Register_Diabetes.htm

References

[1] Carstensen, B. et al.: The Danish National Diabetes Register: Trends in Incidence, Prevalence and Mortality. Diabetologia. 51(12), 2187–2196 (2008). doi: 10.1007/s00125-008-1156-z

[2] Hallgren Elfgren, I.M., Grodzinsky, E., Törnvall, E.: The Swedish National Diabetes Register in clinical practice and evaluation in primary health care. Prim. Health Care Res. Dev. 17(6), pp. 549-558 (2016). doi:10.1017/S1463423616000098

[3] Cooper, J.G., Thue, G., Claudi, T., Løvaas, K., Carlsen, S., Sandberg, S.: The Norwegian Diabetes Register for Adults – an Overview of the First Years. Norsk Epidemiologi. 23 (1), pp. 29-34 (2013)

[4] O'Mullane, M., McHugh, S., Bradley, C.P.: Informing the Development of a National Diabetes Register in Ireland: a Literature Review of the Impact of Patient Registration on Diabetes Care. Inform. Primary Care. 18 (3), pp. 157-68 (2010)

[5] Hallgren Elfgren, I.M., Törnvall, E., Grodzinsky, E.: The Process of Implementation of the Diabetes Register in Primary Health Care. Int. J. of Qual. Health Care. 24 (4), pp. 419-424 (Aug 2012)

[6] Meystre, S., Savova, G., Kipper-Schuler, K., Hurdle, J.F.: Extracting Information from Textual Documents in the Electronic Health Record: A Review of Recent Research. IMIA Yearbook of Medical Informatics, pp. 138-154 (2008)

[7] UMLS, the Unified Medical Language System. https://www.nlm.nih.gov/research/umls/

[8] Denny, J.C., Irani, P.R., Wehbe, F.H., Smithers, J.D., Spickard, A.: The KnowledgeMap Project: Development of a Concept-Based Medical School Curriculum Database. In: AMIA AnnuSymp Proc., pp. 195-199 (2003)

[9] Chapman, W., Bridewell, W., Hanbury, P., Cooper, G.F., Buchanan, B.: A Simple Algorithm for Identifying Negated Findings and Diseases in Discharge Summaries. Univ. of Pittsburgh (2002)

[10] Gindl, S.: Negation Detection in Automated Medical Applications. TUW (2006)

[11] HITEx Manual: https://www.i2b2.org/software/ projects/hitex/hitex_manual.html

[12] Liao, K.P., Cai, T., Savova, G.K., Murphy, S.N., Karlson, E.W., Ananthakrishnan, A.N., Gainer, V.S. et al.: Development of Phenotype Algorithms Using Electronic Medical Records and Incorporating Natural Language Processing. British Med. J., 350 (1): h1885 (2015)

[13] Boytcheva, S.: Shallow Medication Extraction from Hospital Patient Records. Studies in Health Technology and Informatics, 166, pp. 119-128. IOS Press (2011)

[14] Tcharaktchiev, D., Angelova, G., Boytcheva, S., Angelov, Z., Zacharieva, S.: Completion of Structured Patient Descriptions by Semantic Mining. Studies in Health Technology and Informatics, 166, pp. 260-269. IOS Press (2011). doi: 10.3233/978-1-60750-740-6-260

[15] Laney, D.: 3D Data Management: Controlling Data Volume, Velocity, and Variety. META Group Research Note, 6, 10 (2001). https://blogs. gartner.com/doug-laney/files/2012/ 01/ad949-3D-Data-Management-Controlling-Data-Volume-Velocity-and-Variety.pdf

[16] Top 238 Business Analytics Tools. Predictive Analytics Magazine (Feb 2012). http://www.predictiveanalyticstoday.com/top-business-intelligence-tools/

[17] Angelova, G., Nikolova, I., Angelov, Zh.: Embedding Language Technologies in a Data Analytics Tool. Advances in Bulgarian Sciences, pp. 29-42. National Centre for Information and Documentation (2016). ISSN: 1314-3565

[18] Nasreen, S., Azam, M.A., Shehzad, K., Naeem, U., Ghazanfar, M.A.: Frequent Pattern Mining Algorithms for Finding Associated Frequent Patterns for Data Streams: A Survey. Procedia Computer Science, 37, pp. 109-116 (2014)

[19] Boytcheva, S., Angelova, G., Angelov, Z., Tcharaktchiev, D.: Mining Comorbidity Patterns Using Retrospective Analysis of Big Collection of Outpatient Records. Health InfSci Syst. J., Springer (2017). ISSN: 2047-2501 (*to appear*)

[20] Tcharaktchiev, D., Zacharieva, S., Angelova, G., Boytcheva, S. et al.: Building a Bulgarian National Registry of Patients with Diabetes Mellitus. J. of Social Medicine, 2, pp. 19-21 (2015) (*in Bulgarian*)

[21] Boytcheva, S., Angelova, G., Angelov, Z., Tcharaktchiev, D.: Text Mining and Big Data Analytics for Retrospective Analysis of Clinical Texts from Outpatient Care. Cybernetics and Information Technologies, 15 (4), pp. 58-77 (2015). doi: 10.1515/cait-2015-0055

[22] Rabatel, J., Bringay, S., Poncelet, P.: Mining Sequential Patterns: A Context-Aware Approach. Advances in Knowledge Discovery and Management, pp. 23-41. Springer (2013)

[23] Ziembiński, R.Z.: Accuracy of generalized context patterns in the context based sequential patterns mining. Control and Cybernetics, 40, pp. 585-603 (2011)

[24] Yu, H.F., Hsieh, C.J., Chang, K.W., Lin, C.J.: Large Linear Classification when Data Cannot Fit in Memory. ACM Transactions on Knowledge Discovery from Data (TKDD), 5 (4), 23 (2012)

Система информационного поиска на основе тематических моделей

© М.Д. Филин © Т.Ю. Грацианова

Московский государственный университет имени М.В. Ломоносова,
Москва, Россия

maxapple@yandex.ru tgratsianova@cs.msu.su

Аннотация. Описана реализация системы поиска близких по смыслу научных статей, основанная на концепции разведочного поиска. Основной идеей такого поиска является использование тематических моделей, с помощью которых можно представить рассматриваемый корпус документов в векторном пространстве тем для последующего извлечения информации о связях документов между собой. Рассмотрены способы улучшения моделей с помощью аддитивной регуляризации, а также построение визуализации моделей.

Ключевые слова: тематическое моделирование, информационный поиск, аддитивная регуляризация, визуализация, распределение Дирихле, база данных.

Information Retrieval System Based on Topic Models

© M.D. Filin © T.Yu. Gratsianova

Lomonosov Moscow State University,
Moscow, Russia

maxapple@yandex.ru tgratsianova@cs.msu.su

Abstract. The paper describes the implementation of a search system of related scientific articles based on the concept of exploration. The main idea of such a search is the use of thematic models, with the help of which it is possible to present the corpus of documents in a vector space of topics for the subsequent extraction of information about the links of documents among themselves. Discusses ways to improve the models by using the additive of regularization, as well as the construction of visualization models.

Keywords: topic modeling, information retrieval, additive regularization, visualization, Dirichlet distribution, database.

1 Введение

Способы поиска документов, принятые в современных поисковых системах, позволяют осуществлять поиск по ключевым словам. Такой поиск удобен, если пользователь может правильно сформировать поисковый запрос и знает терминологию предметной области. Однако, если пользователь пытается разобраться в новой для него сфере, то он может не знать, как она устроена и в каком направлении вести поиск. Хотелось бы иметь систему поиска, с помощью которой было бы легче изучать незнакомые научные направления, следить за тенденциями и иметь представление о взаимосвязях различных сфер исследований [3].

Для построения такой системы необходимо выбрать исходные данные, разработать процесс их обработки, рассмотреть различные методы построения моделей и выбрать наилучшие. Также важно создать веб-сервис для удобного взаимодействия пользователей с системой.

Труды XIX Международной конференции «Аналитика и управление данными в областях с интенсивным использованием данных» (DAMDID/ RCDL'2017), Москва, Россия, 10–13 октября 2017 года

2 Исходные данные

В качестве исходных данных для построения системы поиска была выбрана коллекция статей портала archive.org. Портал предоставляет бесплатный доступ к различным материалам, в частности, научным статьям. Все данные хранятся в облачном хранилище Amazon S3. Для формирования корпуса текстовых документов был автоматизирован процесс загрузки и обработки данных, а также создан модуль, собирающий метаинформацию для обработанных документов, включающую сведения об авторах, времени публикации, разделе науки, к которому относится документ, и др.

В процессе обработки каждый документ представляется с помощью подхода *bag of words* (не учитывается порядок следования слов в тексте), проводится *стемминг* слов (нахождение неизменяемой части слова) и удаление *стоп-слов* (слов, встречающихся почти в каждом документе) с помощью средств библиотек *nltk* и *Textblob*. Далее

обработанные данные были преобразованы в форматы *Vowpal Wabbit* и *UCI Bag-of-words* для последующего использования при моделировании. В качестве тестового набора было обработано около 15000 статей.

3 Тематическое моделирование

Для анализа корпуса текстов был выбран подход, использующий вероятностное тематическое моделирование, – современный инструмент, предназначенный для выявления тематики коллекций документов [2]. Тематическая модель задает отношение между темами и документами в корпусе. Для каждого документа определено дискретное вероятностное распределение $\theta_d = p(t|d)$ его слов по темам, а тема представляется в виде распределения $\varphi_t = p(w|t)$ слов из фиксированного словаря [3]. Задачу тематического моделирования можно представить в виде задачи матричного разложения: мы хотим разложить исходную матрицу $F = p(w|d)$ уникальных слов на документы в произведение матриц меньшего размера $\Phi = \mathrm{p(w|t)}$ и $\Theta = (t|d)$ с промежуточной размерностью, равной числу тем. Далее, если обозначить множество уникальных слов, встречаемых в текстах, как W, множество документов как D, количество слов w в документе d как n_{dw}, а множество всех встречаемых тем как T, то для оценивания параметров модели $\varphi_{wt} = p(w|t)$ и $\theta_{td} = p(t|d)$ необходимо решить задачу максимизации логарифмированного правдоподобия

$$L(\Phi, \Theta) = \sum_{d \in D} \sum_{w \in d} n_{dw} \log \sum_{t \in T} \varphi_{wt} \theta_{td} \rightarrow \max_{\Phi, \Theta}.$$

Количество тем, встречающихся в документах, намного меньше количества уникальных слов в коллекции, поэтому темы позволяют представить документ в виде вектора в пространстве меньшей размерности, равной числу тем, вместо представления в пространстве слов. В результате документ имеет меньшее число компонент, что позволяет быстрее и эффективнее его обрабатывать.

4 Построение тематических моделей

Для построения моделей используем библиотеки gensim и bigARTM.

Методы тематического моделирования библиотеки gensim основаны на алгоритме латентного размещения Дирихле (Latent Dirichlet Allocation, LDA), который предполагает, что столбцы θ_d и φ_t являются случайными векторами, порождаемыми распределениями Дирихле [1]. При построении моделей можно изменять значения параметров распределения Дирихле и задавать количество тем в корпусе. Путем варьирования значений параметров были найдены оптимальные настройки для моделирования. Для оценивания моделей использовалась *перплексия* (perplexity) – внутренний критерий, определяемый через логарифм правдоподобия:

$\mathcal{P}(D, p) = \exp(-\frac{1}{n} \sum_{d \in D} \sum_{w \in W} n_{dw} \ln p(w|d))$, где $n -$ количество слов в коллекции, $p -$ модель.

4.1 Визуализация

Чтобы понять структуру полученной тематической модели и лучше оценить её качество, можно использовать специальные средства визуализации моделей [4]. Было построено визуальное представление, с помощью которого можно увидеть распределение тем в корпусе документов и слов в темах. Визуализация строится с помощью техники многомерного шкалирования (была использована библиотека *pyLDAvis*), темы обозначаются кругами, схожие по смыслу темы находятся близко друг к другу, а о значимости каждой темы можно судить по величине соответствующего ей круга.

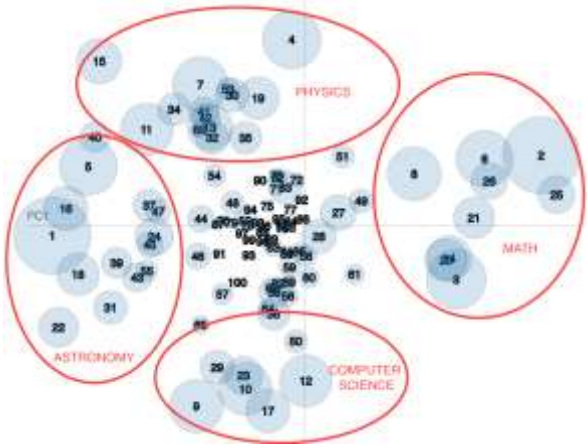


Рисунок 1 Визуализация тематической модели с количеством тем, равным 100

На Рис. 1 видно, что темы, описывающие разные разделы науки, располагаются группами. Темы, имеющие отношение к одному разделу науки, могут быть сгруппированы. Это позволяет судить о том, какими темами представлен тот или иной раздел, какие темы из разных разделов близки между собой и какие темы характеризуют каждый из разделов.

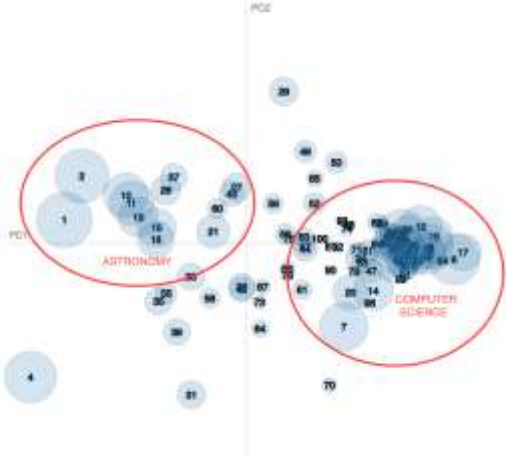


Рисунок 2 Визуализация тематической модели для разделов Computer Science и Astronomy с количеством тем, равным 100

Для демонстрации работы модели рассмотрим ее работу на корпусе, содержащем статьи только по двум направлениям: Computer Science и Astronomy.

Визуализация распределения тем полученной тематической модели хорошо показывает, как образовалось две группы тем (см. Рис. 2). Можно судить о том, что темы, относящиеся к Computer Science, лучше сгруппированы в отличие от тем Astronomy. Такие модели и визуализации можно построить для любых комбинаций выборок из различных разделов науки.

Исследовав тематические модели с помощью визуального представления и экспертного анализа слов, составляющих темы, можно прийти к выводу, что темы, построенные по описанному принципу, могут обладать следующими недостатками:

- плохо интерпретируются (посмотрев на слова, характеризующие тему, нельзя определить предметную область);
- содержат слишком много слов (нет явно выраженного ядра слов с большими вероятностями);
- включают слова общей лексики;
- оказываются слишком похожими между собой.

Также полученная визуализация показала наличие множества фоновых тем в рассматриваемом тестовом наборе. Такие темы включают слова общей лексики и имеют малую смысловую значимость.

Причина этих недостатков заключается в том, что задача построения вероятностной тематической модели по коллекции документов имеет бесконечно много решений, малая часть которых интерпретируема.

4.2 Улучшение моделей

Для получения более качественных моделей применен подход, называемый аддитивной регуляризацией [5]. Основная идея подхода состоит в выборе дополнительных критериев оптимальности, учитывающих специфические требования решаемой задачи, называемых регуляризаторами R, которые учитываются при оптимизации:

$$L(\Phi, \Theta) + R(\Phi, \Theta) \rightarrow \max_{\Phi,\Theta}.$$

Таблица 1 Сравнение моделей PLSA (без регуляризации) и bigARTM для тестового набора документов

	PLSA	bigARTM
Perplexity	2483	2326
Sparsity Φ	0.077	0.985
Sparsity Θ	0.000	0.001
Kernel contrast	0.666	0.738
Kernel purity	0.097	0.210

Библиотека bigARTM позволяет использовать различные регуляризаторы и их комбинации. Также она значительно быстрее, чем gensim, поскольку основные модули написаны на языке C++. Для тестового корпуса были построены модели с помощью данной библиотеки. В качестве дополнительных критериев были добавлены регуляризаторы декореллирования и разреживания тем. В Таблице 1 приведено сравнение моделей.

Для сравнения моделей помимо перплексии использованы следующие оценки:

- Sparsity – разреженность матриц, которая предполагает, что чем выше разреженность, т. е. значительная часть вероятностей φ_{wt} и θ_{td} равна нулю, тем лучше выполняется предположение что каждый документ и каждое слово связано с небольшим числом тем;
- Kernel contrast – показывает, насколько ядра тем отличаются друг от друга. Ядро – множество слов из темы с наибольшими вероятностями, характеризующими тему. Тематическая модель является более качественной, когда темы хорошо различаются;
- Kernel purity – определяет чистоту темы, которая определяется как сумма вероятностей слов, входящих в ядро темы.

5 Структурирование данных

Данные, полученные на этапе сбора информации о документах, необходимо формализовать и преобразовать в наиболее удобное представление для дальнейшего анализа. В качестве реляционной базы данных была использована MySQL. Чтобы структурировать данные о документах, были созданы шесть таблиц, связанные между собой отношениями и приведенные к третьей нормальной форме: documents, authors, tags, topics, sections, comments. Все новые поступающие данные могут быть добавлены в базу данных и хорошо структурированы.

Реализованные процессы загрузки, обработки и хранения данных соответствуют парадигме ETL управления данных, которая включает в себя следующие этапы:

- извлечение данных из внешних источников;
- трансформация и очистка данных;
- структурированное хранение данных.

В итоге полученные данные можно использовать в качестве тестовых данных для многих задач машинного обучения, таких как классификация, кластеризация и др.

6 Организация работы веб-сервиса

Одной из задач при реализации системы поиска являлось создание веб-сервиса, с помощью которого конечный пользователь смог бы осуществлять поиск научных статей и получать информацию о различных параметрах корпуса в целом и каждого документа в отдельности. Для функционирования веб-сервиса был выбран фреймфорк Django, использующий концепцию MVC (Model-View-Controller), позволяющую разделить общую архитектуру на отдельные части: данные, логику и визуализацию.

Разработанная программа предоставляет пользователю возможность совершать поиск тематически похожих документов с помощью двух видов запросов – короткой текстовой строки и целого текстового документа, загружаемого из файловой системы пользователя. В ранжированном списке документов, полученном в ответ на запрос, для

каждого документа выводятся:

- название;
- краткое описание (сниппет) с выделенными цветом словами, соответствующими темам, встречаемым в документе;
- круговая диаграмма распределения тем;
- вес (расстояние между вектором тем запроса t^q и вектором тем документа t^d в соответствии с используемой метрикой $pdist = \sum_{i=0}^{T}(t_i^q - t_i^d)^2$, где $T -$ количество тем в корпусе);
- уникальный идентификатор документа на портале archive.org;
- метаинформация (дата публикации, список авторов и др.).

Рисунок 3 Архитектура программной системы

6.1 Тестирование системы

Чтобы проверить работоспособность системы на базовом уровне, был использован следующий способ. Для каждого документа коллекции создавался запрос на нахождение близких по смыслу. Система должна выдавать сами документы, заданные в качестве запроса, поскольку любой документ ближе всего по смыслу к самому себе. Дальше задача усложнялась: в качестве запроса выступала лишь часть документа. Поскольку используется модель *Bag-of-words*, то из документа можно было изъять какую-то часть составляющих его слов и, используя оставшиеся, выполнить поиск. В итоге, если оставлять 40% от документа, то в 95% случаев после выполнения запроса нужный документ находился в пятерке самых релевантных.

Для более точного тестирования необходимо привлекать асессоров.

7 Заключение

Количество документов, в том числе и научных статей, в современном мире растет, и становятся необходимы специализированные системы поиска информации. В работе рассмотрена возможность применения тематического моделирования для поиска близких по смыслу документов и реализована программная система, которая соответствует концепции разведочного поиска. Запущен сервер в тестовом режиме для анализа работы системы на небольшой части коллекции статей портала *archive.org*.

В дальнейшем планируется усовершенствовать систему, чтобы она помогала пользователям следить за трендами науки, самыми популярными направлениями, новыми статьями и могла рекомендовать пользователям потенциально значимую для них информацию.

Также необходимо обработать все данные портала (порядка 1,5 млн статей), чтобы работать со всей доступной информацией.

Литература

[1] Blei, D., A. Ng., Jordan, M.: Latent Dirichlet Allocation. J. of Machine Learning Research, 3, pp. 993-1022 (2003)

[2] Blei, D.M.: Probabilistic Topic Models. Communications of the ACM, 55 (4), pp. 77-84 (2012)

[3] Ryen, W. White and Resa A. Roth (2009). Exploratory Search: Beyond the Query-Response Paradigm, San Rafael, CA: Morgan and Claypool

[4] Sievert, C., Shirley, K.: LDAvis: A Method for Visualisation and Interpreting Topics. Workshop on Interactive Language Learning, Visualization, and Interfaces (2014)

[5] Vorontsov, K.V., Potapenko, A.A.: Additive Regularization of Topic Models. Machine Learning, Special Issue on Data Analysis and Intelligent Optimization (2014)

Анализ гуманитарных текстов 2

Text analysis in humanities 2

О влиянии семантики на точность определения парафраз в русскоязычных текстах

© К.К. Боярский[1] © Е.А. Каневский[2]

[1] Санкт-Петербургский национальный исследовательский университет информационных технологий, механики и оптики,
[2]Санкт-Петербургский экономико-математический институт РАН,
Санкт-Петербург

boyarin9@yandex.ru kanev@emi.nw.ru

Аннотация. Статья посвящена идентификации парафраз в русскоязычных текстах. В качестве инструмента для решения этой проблемы предлагается использовать семантико-синтаксический парсер SemSin и семантический классификатор. Проанализированы несколько вариантов определения близости пар предложений: по словам, по леммам, по классам, по семантически связанным концептам, по предикатным группам. Обсуждены преимущества и недостатки этих методов. Показано, что при увеличении глубины использования семантической информации качество идентификации парафраз повышается. Однако включение в анализ предикатных групп, определяемых по дереву зависимостей, может даже привести к ухудшению качества идентификации вследствие увеличения числа ложноположительных решений.

Ключевые слова: парафразы, семантические словари, леммы, классификатор, семантические классы, парсинг, синонимия.

Effect of Semantic Parsing Depth on the Identification of Paraphrases in Russian Texts

© K. Boyarsky[1] © E. Kanevsky[2]

[1]ITMO University,
[2]EMI RAS,
St Petersburg, Russia

boyarin9@yandex.ru kanev@emi.nw.ru

Abstract. As a tool to solve the problem of identification of paraphrases in Russian texts, the paper proposes the semantic-syntactic parser SemSin and a semantic classifier. Several alternative methods for evaluating the similarity of sentence pairs – by words, by lemmas, by classes, by semantically related concepts, by predicate groups – have been analyzed. Advantages and drawbacks of the methods are discussed. The paraphrase identification quality has been shown to rise with increasing depth of using the semantic information. Yet, complementing the analysis with predicate groups, identified by the dependency tree, may even cause the identification to degrade due to the growing number of false positive decisions.

Keywords: Russian texts, paraphrases, semantic dictionary, lemmas, classifier, classes, semantic-syntactic parsing, synonymy.

1 Введение

В последнее время значительный интерес исследователей, работающих в области поиска информации, привлекает проблема выявление парафраз.

В англоязычной литературе существует большое количество работ по идентификации парафраз с привлечением различной лексической, синтаксической и семантической техники [3, 17]. В большинстве способов использовалось обучение, проводились токенизация, определение частей речи и обработка только существительных и глаголов [6]. Использовалось также придание различного веса словам с учетом их грамматической роли в предложениях. Corley and Mihalcea [4] использовали измерения семантической близости текстов с помощью WordNet [8]. При этом семантическое сходство слов измерялась только для глаголов и существительных, а сравнение наречий, прилагательных и количественных числительных

Труды XIX Международной конференции «Аналитика и управление данными в областях с интенсивным использованием данных» (DAMDID/ RCDL'2017), Москва, Россия, 10–13 октября 2017 года

проводилось лексически. Было показано, что такой метод значительно точнее, чем простое лексическое сравнение. Близкая техника, основанная на семантической информации WordNet, использовалась в [5]. Pershina [13] дополнительно для идентификации парафраз использовала базу идиом. Лучшие результаты по идентификации парафраз в англоязычных текстах дают F-меру около 82%.

Для русскоязычных текстов работы по идентификации парафраз весьма немногочисленны [18]. Значительные сложности возникают в связи с особенностями русского языка, который отличается свободным порядком слов и богатой морфологией [16]. Целью настоящей работы являются исследование эффективности различных вариантов анализа парафраз в русскоязычных текстах и определение оптимальной степени использования семантической информации.

Данная работа проводилась в рамках конкурса по идентификации русских парафраз, представленных в корпусе paraphraser.ru [10]. В рамках этого конкурса предлагалось два варианта анализа пар новостных заголовков: разделение на две группы (парафразы и не парафразы) и на три группы, с выделением дополнительно группы нечетких парафраз. По представленным данным оказалось, что первый вариант анализа обладает существенно большей точностью. В значительной мере это связано с субъективностью выделения нечетких парафраз. Поэтому было выбрано разделение на две группы.

Для определения того, являются ли два предложения парафразами, т. е. одинаков ли их смысл для носителей языка, необходимо ввести числовую меру сходства. В качестве такой меры мы использовали коэффициент Жаккара J. Если два предложения A и B содержат соответственно $n(A)$ и $n(B)$ лексических единиц, то

$$J = \frac{n(A) \cap n(B)}{n(A) \cup n(B)}.$$

В соответствии с этим критерием мера сходства определяется как отношение числа совпадающих единиц к общему числу различных единиц.

Следует решить два вопроса: что считать лексической единицей, подлежащей сравнению, и каков критерий собственно сравнения. Для сравнения нами были выбраны четыре варианта.

Вариант 1. В качестве единицы для сравнения принимается слово или иная непрерывная последовательность знаков. Предполагается, что эта последовательность достаточно длинная, так что одно- и двухбуквенные слова не учитываются.

Преимущества. Не требуется каких-либо средств для разбора текста. Выявление совпадающей группы знаков с большой вероятностью означает наличие однокоренных слов, близких по смыслу. Если же такая группа представляет отдельное слово, то возможно, что это слово выполняет одну и ту же функцию в обоих высказываниях.

Недостатки. Похожие фрагменты с большой вероятностью включают в себя служебные слова. Их совпадение часто носит случайный характер и не отражает смысла фрагмента текста (хотя, например, совпадение отрицания *не* может служить для определения меры близости). Кроме того, поскольку русский язык относится к синтетическому типу, то даже очень небольшая перефразировка, совершенно не меняющая смысл высказывания, приводит к изменению словоформ. Это существенно снижает точность анализа.

Вариант 2. В качестве единицы для сравнения вместо словоформы принимается нормализованная форма слова (лемма). Исключаются служебные слова. Производится настройка парсера на предметную область путем исключения омонимичных словоформ, не принадлежащих к данной предметной области. Например, словоформе *белку* соответствуют две леммы: *белок* и *белка*. В тексте по биологии будет выбрана только первая из них.

Преимущества. Устраняются сложности, связанные с богатством словоформ каждой леммы. Повышается независимость анализа от конкретного строя предложения (*человек, который смеется* vs *смеющийся человек*).

Недостатки. Требуется инструмент для морфологического анализа. Достаточно часто (примерно в 7% случаев) лемма по словоформе определяется неоднозначно, т. е. возникает проблема омонимии. Остаются вопросы, связанные со служебными словами.

Вариант 3. При сравнении учитывается семантика слов.

Преимущества. Слова сравниваются не столько по написанию, сколько по смыслу, что, в принципе, должно повысить точность определения парафраз.

Недостатки. Сложность и неоднозначность семантического анализа. Необходимость использования семантических словарей. Если в английском языке эту функцию выполняет semantic web, то для русского языка соответствующие инструменты слабо развиты, отсутствует принятый в качестве стандарта де-факто семантический словарь.

Вариант 4. Сравнение производится с учетом полного дерева разбора.

Преимущества. Возможность анализа сходства неконтактно стоящих групп слов, выделения смысловых блоков, описывающих термины предметной области.

Недостатки. Повышение вероятности ошибок парсера. Недостаточная проработанность вопроса о выделении контекстных блоков. Чувствительность метода к замене одних оборотов другими, например, причастного оборота придаточным предложением.

Из русскоязычных парсеров, способных проводить глубинный анализ предложения, наиболее известны ЭТАП-3 [12], а также парсеры фирм Яndex [2], Abbyy [1], «Диктум» [9]. Два первых парсера работают с помощью системы правил, два других используют свою собственную технологию. Все парсеры тем или иным образом используют словари.

В данной работе использовался семантико-синтаксический парсер SemSin [22]. Это система, сочетающая в себе функции лемматизатора,

синтаксического и семантического анализатора. Парсер включает в свой состав словари, классификатор, блок морфологического анализа, предсинтаксический модуль [21] и набор продукционных правил [23,24].

В процессе анализа предложения одновременно выполняются снятие грамматической и частеречной омонимии, сегментация предложения и построение синтаксического дерева зависимостей.

Полученное дерево содержит максимально полную информацию о предложении. Эта информация может в дальнейшем служить основой для решения самых разных задач: выявления терминов [15], классификации текстов [7] и т. д. В данной работе обсуждается, какая именно информация полезна для определения близости смысла предложений, т. е. для идентификации парафраз.

2 Анализ текста

Основное внимание было уделено анализу различных способов описания семантики и включения ее в процедуру идентификации парафраз. Все примеры и экспертные оценки брались из корпуса русских парафраз [19].

2.1 Лемматизация

Как было показано в [14], в русском языке (в отличие от английского и французского) в задачах кластеризации текстов наибольшей дифференцирующей силой обладают существительные. Однако для идентификации парафраз выделение только существительных слишком грубо и не может охватить всех тонкостей смысла. Поэтому для сравнения оставлялись существительные, прилагательные, глаголы и отглагольные формы (причастия, деепричастия) и числительные. Примеры влияния лемматизации на величину коэффициента Жаккара приведены в табл. 1.

Таблица 1 Влияние лемматизации

	Предложения	Словоформы	Леммы
1	NI опубликовал список самого опасного вооружения флота России В США опубликован топ-5 самых опасных вооружений ВМФ России	0.067	0.455
2	Путин впервые объявил минуту молчания на параде Победы Пан Ги Мун поблагодарил Путина за организацию Парада Победы	0.067	0.250

Отметим, что лемматизация увеличивает степень согласованности предложений как в случае, когда они являются парафразами (пример 1), так и когда совпадение слов случайно, а совпадения смысла нет (пример 2). Таким образом, лемматизация безусловно повышает точность сравнения лексики предложений, но недостаточна для сравнения смысла.

2.2 Семантика. Учет классов

Для более точного сравнения предложений был использован семантический классификатор, содержащий около 1700 классов. Его основой является классификатор Тузова [27], ориентированный именно на компьютерный анализ текстов. Дерево классов построено с таким расчетом, чтобы семантические классы имели определенные синтаксические свойства. Например, список действий, которые может производить живое существо, отличается от действий неодушевленных предметов, у разных классов могут быть разные атрибуты и т. д. Кроме того, формат этого классификатора удобен для его автоматического использования.

Наш классификатор отличается, в частности, от классификатора Шведовой [26]. Например, слово *жрец* в обоих случаях классифицируется практически одинаково: как определенный тип профессии человека. В то же время слово *желание* у Шведовой находится в ветви дерева *духовный мир–чувства–*…, в то время как в нашем классификаторе психические явления и чувства рассматриваются как свойства человека. В классификаторе WordNet положение слова *priest* в общем соответствует русским классификаторам, слова *desire* – ближе к классификатору Шведовой.

При разработке классификатора очень трудно, если вообще возможно, определить, на каком ярусе дерева нужно поместить то или иное слово, по какому признаку разделять подклассы, в какой момент прекращать дальнейшее ветвление. Например, *коса* (*scythe*) в WordNet относится к классу режущих инструментов, а в нашем классификаторе и у Шведовой – к классу сельскохозяйственных орудий. При идентификации парафраз принималось, что принадлежность разных слов в первом и втором предложениях одному классу означает близость их смысла.

Часто вместо конкретного слова в тексте появляется его гипероним. Для обнаружения гиперонимов класс определялся с точностью ±1 уровень иерархии. Таким образом обеспечивается совпадение слов, выделенных полужирным шрифтом в следующих примерах.

Лавров подарил Керри ***помидоры****… vs Лавров подарил Керри* ***овощи****…*

Жертвами взрыва … стали не менее трех ***человек*** *vs Жертвой взрыва… стал* ***гражданин*** *Великобритании.*

Исключение составляют имена собственные, поскольку, например, все названия городов относятся к одному классу, также к одному классу относятся все фамилии людей.

Сравнением классов во многих случаях перекрываются отношения синонимии. Однако следует иметь в виду, что из-за морфологических

особенностей в русском языке значительно меньше совпадений словоформ у различных частей речи. Так? в английском *gold* это и существительное, и прилагательное, в русском – существительное *золото*, прилагательное *золотой*. При определении синонимии в русскоязычном RusNet [20] подразумевается совпадение частей речи у синонимичных слов. Сравнение «по классам» в этом смысле шире и позволяет делать заключение о близости смысла даже при существенной перефразировке:

Турецкий сухогруз подвергся ***обстрелу****… vs Турецкий сухогруз* ***обстреляли****.*

Несомненно, вопрос о том, являются ли два существительных, относящихся к одному классу, синонимами, неоднозначен. Например, слова *веревка*, *бечевка* легко могут взаимозаменяться в тексте, а слова *помидор*, *огурец* – нет. Тем не менее, анализ показывает, что близость классов чаще всего означает близость смысла.

2.3 Семантика. Синонимия и семантические гнезда

В ряде случаев сравнения по классам оказывается недостаточно. Известна, например, «теннисная проблема» [25], заключающаяся в том, что слова, относящиеся к одной предметной области, часто находятся в совершенно разных ветвях классификатора, что затрудняет не только идентификацию парафраз, но и решение задач классификации и кластеризации текстов. Наш словарь содержит дополнительную информацию о семантической близости слов. Будем называть группу семантически связанных слов семантическим гнездом. Фактически появление дополнительных связей означает превращение дерева классов в семантическую сеть. Имеется несколько ситуаций, приводящих к образованию семантических гнезд.

- Производные слова

Все лексемы в словаре делятся на базовые и производные, причем под производными подразумеваются не только слова, однокоренные с базовыми [27]. Базовых лексем несколько меньше половины (около 83000). Из них примерно 20000 образуют семантические гнезда, к которым относятся производные слова с близким смыслом. Так, к гнезду базового слова *чувство* принадлежит более 100 производных слов, среди которых есть существительные (*нечувствительность*, *аналгезия*), прилагательные (*чувствительный*, *душещипательный*), глаголы (*чувствовать*, *обуревать*). В некоторых случаях семантический класс производного слова совпадает с классом базового, в других – отличается. Например, базовое слово *сигнал* относится к подклассу ветви семантического дерева *информация*. Производные от него слова: *сигнальный, сигнализация, сигнализировать* имеют тот же класс, а слово *сигнальщик* – человек с определенным родом занятий – совсем другой. Принадлежность производных слов к общему гнезду добавляет около 7500 связей в семантическую сеть (помимо связей класс – подкласс).

- Географические названия

Как указывалось выше, имена собственные сравниваются только по леммам. Но одна и та же страна может называться по-разному, и это необходимо учитывать при анализе парафраз. Часто в качестве названия страны используется аббревиатура. Поэтому в словарь были внесены дополнительные сведения о тождественной синонимии слов и выражений.

В ***Британии*** *палата общин одобрила однополые браки.*

Палата общин ***Великобритании*** *одобрила однополые браки.*

США просят ***РФ*** *немедленно отменить запрет на ввоз мяса.*

США призвали ***Россию*** *немедленно снять запрет на импорт мяса.*

КНДР *готовится нанести ракетный удар по США.*

Северная Корея *пригрозила ракетным ударом по США.*

Неким аналогом отношений класс – подкласс для географических названий является информация о принадлежности населенного пункта к региону и стране. Такая информация была добавлена в словарь, хотя и в довольно ограниченном размере. Например, указано, что *Дамаск* – столица *Сирии*, а город *Сент-Луис* находится в штате *Миссури* страны *США*:

Неизвестный открыл огонь в бизнес-школе штата ***Миссури****.*

В ***Сент-Луисе*** *преступник открыл огонь в бизнес-школе.*

- Характеристики людей и организаций

Следующей группой слов, для которых в словарь были внесены дополнительные связи, приводящие к образованию семантических гнезд, являются характеристики людей по национальности и месту жительства. Необходимо не только знать, что *сибиряк* – название человека по месту жительства, но и что это место – именно *Сибирь*. Таким образом, каждое такое слово «прикреплено» к соответствующей стране, региону или городу:

Американец *выиграл в лотерею за 100 евро картину П. Пикассо.*

Житель ***Пенсильвании*** *выиграл в лотерею картину Пикассо.*

Для некоторых часто встречающихся имен высокопоставленных деятелей в качестве составляющих семантического гнезда указаны их должности и страны:

Синдзо ***Абэ*** *в письме* ***президенту РФ*** *объяснил, почему не приедет на 9 мая.*

В письме ***Путину японский премьер*** *объяснил причины отказа приехать в Москву 9 Мая.*

США *приостановили поставку истребителей в Египет.*

Обама *приостановил поставки истребителей F-16 в Египет.*

Еще одну группу семантических гнезд составляют связи по принадлежности определенных социальных групп к организациям и т. д. *Коммунист* – наименование человека не просто по принадлежности к какой-то общественной организации, а именно к компартии, а *хоккеист* среди всех видов спорта имеет отношение только к хоккею:

Депутаты от ***КПРФ*** *попросили Путина взять под защиту гималайского медведя.*

Коммунисты *попросили Путина защитить гималайских медведей.*

Сборная России по ***хоккею*** *проиграла финнам и во втором матче Евротура.*

Российские ***хоккеисты*** *проиграли на Евротуре четыре матча подряд.*

- Идиомы

Отдельную группу семантических гнезд образуют связи устойчивых выражений и идиом с их семантическими аналогами:

В Красноярском крае ***исчез*** *заместитель прокурора.*

В Красноярском крае ***пропал без вести*** *помощник прокурора района.*

Ушел из жизни *Уго Чавес.*

Умер *Уго Чавес.*

Рассмотрим следующий пример.

Оппозиция ФРГ угрожает правительству судом из-за шпионского скандала.

Немецкая оппозиция пригрозила правительству иском из-за скандала с BND.

Если сравнивать эти предложения только по леммам, то имеются три совпадения (*оппозиция*, *правительство*, *скандал*) с коэффициентом Жаккара *J*=0.27. При учете классов получаем совпадение для глаголов угрожать и пригрозить, становится *J*=0.40. Название *ФРГ* входит в семантическое гнездо слова *Германия*, в это же гнездо входит слово *немецкий* (производное от базового слова *немец*, которое, в свою очередь связано со словом *Германия*). Получаем *J*=0.55. Наконец, лексемы *суд* и *иск* тоже входят в общее гнездо. Окончательно получаем *J*=0.75, что вполне отражает смысловую близость этих предложений. На данный момент число семантических связей таких типов в словаре около 12 тыс.

2.4 Дерево зависимостей

Нами была сделана попытка использовать построенное парсером дерево зависимостей для уточнения сравнения смысла предложений. Было выдвинуто предположение, что совпадение субъекта действия и собственно предиката повышает вероятность того, что рассматриваемые предложения являются парафразами. В этом случае при расчете коэффициента Жаккара совпадение лемм учитывалось с весовым коэффициентом 1.5.

Однако на практике оказалось, что у пар предложений, обладающих указанным свойством, коэффициент Жаккара обычно и так уже велик. Поэтому его незначительное увеличение сравнительно редко переводит эту пару из разряда «не парафраза» в разряд «парафраза». Но эти редкие переходы тоже представляют определенный интерес. В табл. 2 приведены значения коэффициентов Жаккара при учете совпадений лемм, классов и семантических гнезд (LCS) и дополнительно предикатных пар (LCSP).

Таблица 2 Влияние совпадения пары субъект–предикат на величину коэффициента Жаккара

	Предложения	LCS	LCSP	Комментарий
1	МИД Чехии: дипломат получил выговор за высказывание о пожаре в Одессе Оправдавший сожжение людей в Одессе чешский дипломат получил выговор	0.33	0.50	Верный результат, предложения идентичны по смыслу
2	Кобзон назвал результат России на Евровидении очень достойным Иосиф Кобзон назвал второе место Полины Гагариной достойным	0.27	0.45	Верный результат, предложения идентичны по смыслу
3	Лужков назвал свою ферму примером для российского правительства Лужков назвал российскую экономику «антинародной»	0.37	0.62	Ошибочный рост коэффициента совпадения, связанный с игнорированием прямого дополнения к переходному глаголу
4	МВД насчитало 200 тыс. участников празднования Дня Победы в Севастополе МВД насчитало 250 тыс. участников акции «Бессмертный полк» в Москве	0.28	0.48	Ошибочный рост коэффициента совпадения, связанный с игнорированием различия в месте действия
6	Эксперты из России и Белоруссии направились с проверкой в Эстонию Российские военные эксперты направились с проверкой в Эстонию	0.71	0.86	В стандарте не считаются парафразами. Несомненно, это одно и то же событие, очевидно, ошибка разметки

Таблица показывает, что учет совпадений субъект – предикат в некоторых случаях облегчает нахождение парафраз (примеры 1, 2). Однако это может привести и к ложному повышению степени сходства предложений (примеры 3, 4). В некоторых случаях определение того, является ли пара предложений парафразами, имеет пограничный характер и может интерпретироваться различными экспертами по-разному (пример 5). Иногда возможны просто экспертные ошибки (пример 6). В общем учет совпадений пары субъект – предикат привел к незначительному снижению качества анализа. Однако это изменение лежит в пределах погрешности. Нам представляется перспективным направлением развития работы дальнейшее расширения анализа дерева зависимостей с целью снижения коэффициента сходства у пар предложений, отличающихся, например, по месту действия, и уменьшению числа ложных срабатываний.

3 Результаты

Ниже приведены результаты анализа смысловой близости пар предложений для выявления парафраз, выполненного по следующим схемам.

По словам (W). При сравнении учитывались все слова, включая служебные части речи.

По леммам (L). Учитывались существительные, прилагательные, глаголы, числительные.

По леммам и классам (LC). Дополнительно учитывалось совпадение классов по семантическому дереву.

По леммам, классам и семантическим гнездам (LCS). Дополнительно учитывалось совпадение классов по семантической сети.

Примеры влияния схемы расчета на величину коэффициента Жаккара приведены в табл. 3.

Таблица 3 Коэффициент Жаккара для различных схем подсчета совпадений

№	Предложения	W	L	LC	LCS
1	Крупный пожар вспыхнул на складе на северо-востоке Москвы Крупный пожар в административном здании в центре Москвы потушен	0.230	0.300	0.444	0.444
2	Продажи АвтоВАЗа в России в апреле сократились на 38,3% Продажи «АвтоВАЗа» в России рухнули на треть	0.273	0.500	0.667	0.778
3	СМИ: поздравляя Вакарчука, Кличко ошибся с возрастом и именем юбиляра Кличко перепутал в поздравлении имя и возраст солиста «Океана Эльзы»	0.062	0.230	0.455	0.455
4	Выселен последний экс-депутат, незаконно занимавший жилье в Москве В Москве выселили бывшего депутата Думы из служебной квартиры	0.071	0.181	0.300	0.600
5	Morgan Stanley взял на работу бывшего зампреда Банка России. Бывший глава ФСФР нашел работу в Morgan Stanley	0.200	0.364	0.500	0.500

На этапе обучения системы проводилось пополнение словаря специфическими терминами, а также определение оптимального «уровня отсечки»: с какого значения коэффициента Жаккара считать пару предложений парафразом. Рассчитывались стандартные параметры: аккуратность (A), точность (P), полнота (R) и F-мера (F):

Типичные распределения показаны на рис. 1.

Зависимости имели качественно сходный характер и для остальных схем подсчета. Очевидно, что, снижая уровень отсечки, т. е. относя к парафразам почти все пары предложений, можно добиться сколь угодно высокого значения полноты, а повышая этот уровень, – сколь угодно высокого значения точности. Оценка качества анализа проводилась по параметрам аккуратность и F-мера, для которых оптимальный уровень отсечки оказался в интервале 0.35…0.40.

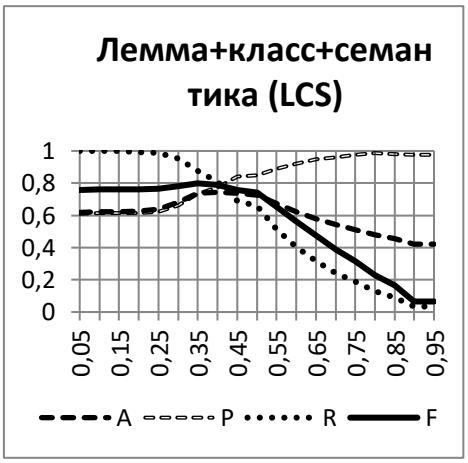


Рисунок 1 Зависимость параметров точности от уровня отсечки

Таблица 4 Аккуратность и F-мера при различных схемах сравнения

Схема	W	L	LC	LCS	LCSP
A	0.692	0.718	0.743	0.744	0.742
F	0.762	0.774	0.784	0.800	0.795

На основании этой таблицы можно сделать следующие выводы.

4 Заключение

Результаты по качеству идентификации парафраз приведены в табл. 4. Сравнение проводилось по разметке Золотого стандарта [11]. Для каждой из рассмотренных схем сравнения приведены лучшие значения аккуратности A и F-меры F.

В большинстве случаев заголовки новостей, относящиеся к одному и тому же событию, лексически очень сходны, и могут быть определены как парафразы любым способом.

Применение методов семантико-синтаксического сравнения (LC) улучшает результаты по сравнению не только с простым посимвольным сравнением, но и со сравнением по леммам.

Увеличение «глубины» семантического анализа за счет перехода от дерева классов к семантической сети (LCS) улучшает качество анализа.

Дополнительный учет совпадений субъект – предикат (LCSP) незначительно ухудшает качество за счет увеличения числа ложных срабатываний.

Заметим, что в статье [18], табл. V, приводятся 15 пар предложений, особенно трудных для анализа парафраз. Три метода, рассматриваемые в этой статье, дают соответственно 6, 6 и 7 ошибок. Предлагаемый нами метод LCS дает только 2 ошибки.

Таким образом, достигнутые показатели качества лежат на уровне лучших результатов конкурса по классификации предложений на две группы: парафразы и не парафразы (у лидеров A=0.7459 и F=0.8078 [10]). Наши показатели также вполне сравнимы с результатами, получаемыми на англоязычных текстах. При классификации на три группы наши результаты существенно ниже из-за сложности выделения группы «сомнительных парафраз».

Преимуществом обсуждаемого метода является то, что в процессе работы используется только словарная информация, так что переобучение системы при смене предметной области не требуется.

Литература

[1] Anisimovich, K.V., Druzhkin, K.Ju., Minlos, F.R., Petrova, M.A., Selegey, V.P., Zuev, K.A.: Syntactic and Semantic Parser Based on ABBYY Compreno Linguistic Technologies. Компьютерная лингвистика и интеллектуальные технологии. По материалам ежегодной Межд. конф. «Диалог». М.: РГГУ, 2 (11(18)), сс. 91-103 (2012)

[2] Antonova, A.A., Misyurev, A.V.: Russian Dependency Parser SyntAutom at the DIALOGUE-2012 Parser Evaluation Task. Компьютерная лингвистика и интеллектуальные технологии. По материалам ежегодной Межд. конф. «Диалог». М.: РГГУ, 2 (11(18)), сс. 104-118 (2012)

[3] Barron-Cedeno, A., Vila, M., Marti, M.A., Rosso, P.: Plagiarism Meets Paraphrasing: Insights for the Next Generation in Automatic Plagiarism Detection. Computational Linguistics, 39 (4), pp. 917-947 (2012)

[4] Corley, C., Mihalcea, R.: Measuring the Semantic Similarity of Texts. Proc. of the ACL Workshop on Empirical Modeling of Semantic Equivalence and Entailment, Association for Computational Linguistics Stroudsburg, PA, USA, pp. 13-18 (2005)

[5] Fernando, S., Stevenson, M.: A Semantic Similarity Approach to Paraphrase Detection. Proc. of the Computational Linguistics UK, 11th Annual Research Colloquium (2008)

[6] Finch, A., Hwang, Y.S., Sumita, E.: Using Machine Translation Evaluation Techniques to Determine Sentence-Level Semantic Equivalence. Proc. of the 3rd Int. Workshop on Paraphrasing, pp. 17-24 (2005)

[7] Artemova, G., Boyarsky, K., Gouzйvitch, D., Gusarova, N., Dobrenko, N., Kanevsky, E., Petrova, D.: Text Categorization for Generation of Historical Shipbuilding Ontology. Communications in Computer and Information Science, 468, pp. 1-14 (2014)

[8] http://wordnet.princeton.edu/

[9] http://www.dictum.ru/ru/syntax-analysis/blog

[10] http://www.paraphraser.ru.

[11] http://www.paraphraser.ru/download/get?file_id=5.

[12] Iomdin, L., Petrochenkov, V., Sizov, V., Tsinman, L.: ETAP Parser: State of the Art. Компьютерная лингвистика и интеллектуальные технологии. По материалам ежегодной Межд. конф. «Диалог». М.: РГГУ, 2 (11(18)), сс. 119-131 (2012)

[13] Pershina, M., He, Y. Grishman, R.: Idiom Paraphrases: Seventh Heaven vs Cloud Nine. Proc. of the EMNLP 2015 Workshop on Linking Models of Lexical, Sentential and Discourse-level Semantics, pp. 76-82 (2015)

[14] Avdeeva, N., Artemova, G., Boyarsky, K., Gusarova, N., Dobrenko, N., Kanevsky, E.: Subtopic Segmentation of Scientific Texts: Parametr Optimisation. Communications in Computer and Information Science, 518, pp. 3-15 (2015)

[15] Avdeeva, N., Boyarsky, K., Kanevsky, E.: Extraction of Low-frequent Terms from Domain-specific Texts by Cluster Semantic Analyses. Proc. of the ISMW-FRUCT 2016. Saint-Petersburg, Russia, FRUCT Oy, Finland, pp. 86-89 (2016)

[16] Nivre, J., Boguslavsky, I.M., Iomdin, L.L. Parsing the SynTagRus Treebank of Russian. Proc. of the 22nd Int. Conf. on Computational Linguistics, 1, pp. 641-648. Association for Computational Linguistics (2008)

[17] Pham, N., Bernardi, R., Zhang, Y.Z., Baroni, M.: Sentence Paraphrase Detection: When Determiners and Word Order Make the Difference. Proc. of the Towards a Formal Distributional Semantics Workshop at IWCS, pp. 21-29 (2013)

[18] Pronoza, E., Yagunova, E.: Comparison of Sentence Similarity Measures for Russian Paraphrase Identification. Artificial Intelligence and Natural Language and Information Extraction, Social Media and Web Search FRUCT Conf., pp. 74-82 (2015)

[19] Pronoza, E., Yagunova, E., Pronoza, A.: Construction of a Russian Paraphrase Corpus:

Unsupervised Paraphrase Extraction. Proc. of the 9th Russian Summer School in Information Retrieval, August 24–28, 2015, Saint-Petersburg, Russia (RuSSIR 2015, Young Scientist Conference), Springer CCIS

[20] Азарова, И.В., Митрофанова, О.А., Синопальникова, А.А.: Компьютерный тезаурус русского языка типа WordNet. Компьютерная лингвистика и интеллектуальные технологии. Труды Межд. конф. «Диалог 2003». М.: Наука, сс. 43-50 (2003)

[21] Боярский, К.К., Каневский, Е.А.: Предсинтаксический модуль в анализаторе SemSin. Интернет и современное общество: сб. научных статей. Труды XVI Всерос. объединенной конф. «Интернет и современное общество». СПб.: НИУ ИТМО, сс. 280-286 (2013)

[22] Боярский, К.К., Каневский, Е.А.: Семантико-синтаксический парсер SemSin. Научно-технический вестник информационных технологий, механики и оптики, 15 (5), сс. 869-876 (2015)

[23] Боярский, К.К., Каневский, Е.А.: Система продукционных правил для построения синтаксического дерева предложения. Прикладна лінгвістика та лінгвістичні технології: MegaLing-2011. К.: Довіра, сс. 73-80 (2012)

[24] Боярский, К.К., Каневский, Е.А.: Язык правил для построения синтаксического дерева. Интернет и современное общество: Материалы XIV Всерос. объединенной конф. «Интернет и современное общество». СПб.: ООО «МультиПроджектСистемСервис», сс. 233-237 (2011)

[25] Лукашевич, Н.В.: Тезаурус в задачах информационного поиска. М.: МГУ, 495 с. (2011)

[26] Русский семантический словарь. Толковый словарь, систематизированный по классам слов и значений / Российская академия наук. Ин-т рус. яз. им. В.В. Виноградова; Под общей ред. Н.Ю. Шведовой. М.: «Азбуковник» (1998)

[27] Тузов, В.А.: Компьютерная семантика русского языка. СПб: Изд-во С.-Пб. ун-та, 391 с. (2004)

Recognizing Names in Islam-Related Russian Twitter

© V. Mozharova © N. Loukachevitch

Lomonosov Moscow State University,
Moscow, Russia

valerie.mozharova@gmail.com louk_nat@mail.ru

Abstract. The paper describes an approach to creating a domain-specific tweet collection written by users frequently discussing Islam-related issues in Russian. We use this collection to study specific features of named entity recognition on Twitter. We found that in contrast to tweets collected randomly, our tweet collection contains relatively small number of spelling errors or strange word shortenings. Specific difficulties of our collection for named entity recognition include a large number of Arabic and other Eastern names and frequent use of ALL-CAPS spelling for emphasizing main words in messages. We studied the transfer of NER model trained on a news wire collection to the created tweet collection and approaches to decrease the degradation of the model because of the transfer. We found that for our specialized text collection, the most improvement was based on normalizing of word capitalization. Two-stage approaches to named entity recognition and Word2vec-based clustering were also useful for our task.

Keywords: NER, CRF, Twitter.

1 Introduction

Named entity recognition (NER) is one of the basic natural language processing tasks [17, 20]. Recognition of named entities in texts is used in many other information-processing tasks as relation extraction, entity linking, information retrieval etc. Most studies of NER have been carried out on news collections and shown high quality of named entity extraction. However, the transfer of NER recognizers to other genres of texts demonstrated significant decrease in the performance.

Currently, there is a great interest in information extraction from texts published on social media platforms such as Twitter or Facebook because these platforms can serve as a very useful (fast and/or alternative) source of information [22]. But application of general NER recognizers designed for or trained on news collections can demonstrate the decrease in performance of up to 50% on more in these informal texts [4, 7–9].

Another important direction of social network studies is directed to differences of language and style in specific social media communities [11, 18] or their dependence on social and demographic characteristics of users [12, 21].

In this paper, we consider the transfer of Russian NER recognizer trained on news texts to extracting names from Twitter messages. Our tweet collection is specialized; it is gathered from messages of those users who discuss issues related to Islam in their posts in contrast to other studies where Twitter collections are formed with random sampling of Twitter messages. This allows us to reveal specific features of the tweet language of Islam-oriented and other similar communities. We consider the transfer of CRF-based NER recognizer from a news data to the tweet collection and approaches to decrease the degradation of the model because of the transfer.

Proceedings of the XIX International Conference "Data Analytics and Management in Data Intensive Domains" (DAMDID/RCDL'2017), Moscow, Russia, October 10–13, 2017

2 Related works

2.1 Named Entity Recognition for Twitter

It is known that extraction of names from Twitter messages is much more difficult task than from other genres of text because of their shortness and informal character.

In [7] the authors review the problems and approaches to named entity recognition and entity linking for tweets. They write that the tweet content is noisy because of incorrect spelling, irregular capitalization, and unusual abbreviations. In their experiments, the main sources of mistakes in named entity recognition in tweets were violations in capitalization especially a large number of names written in lower case. They studied automatic normalization of tweets including spelling and capitalization correction and reported that in their investigation the normalization slightly improved the performance in NER for tweets.

In [24] the authors write that due to unreliable capitalization in tweets, common nouns are often misclassified as proper nouns, and vice versa. Some tweets contain all lowercase words (8%), whereas others are in ALL CAPS (0.6%). In addition to differences in vocabulary, the grammar of tweets differs from news text, for example, tweets often start with a verb. In their experiments, the supervised approach was used to predict correct capitalization of words. The set of features included: the fraction of words in the tweet which are capitalized, the fraction which appear in a dictionary of frequently lowercase/capitalized words but are not

lowercase/capitalized in the tweet, the number of times the word 'I' appears lowercase and whether or not the first word in the tweet is capitalized.

To study NER on Twitter performed with several NER systems, [8] use crowdsourcing to annotate tweets for the NER task. They annotate all @user-names as PER (person name). Annotating tweets for their experiments [24] choose not to annotate @usernames mentioned in tweets as entities because it is trivial to identify them using a simple regular expression, and they would only serve to inate the performance statistics.

In [4] the authors study the transfer of their NER model from news texts to tweets. They create a training set consisting of 1000 tweets. They use a baseline NER model based on token and context features (wordform, lemma, capitalization, prefixes and suffixes) and enhance it with two unsupervised representations (Brown clusters and vector representations) based on a large collection of unannotated tweets. Besides, they propose a technique to combine a relatively small Twitter training set and larger newswire training data. They report that two unsupervised representations work together better than alone, and the combination of training sets further improves the performance of their NER system.

2.2 Named Entity Recognition in Russian

In Russian there is a long tradition of engineering approaches to the named entity recognition task [13, 14, 23].

Machine-learning approaches for Russian NER usually employ the CRF machine learning method. In [1] the authors presented the results of the CRF-based method on various tasks, including the named entity recognition. The experiments were carried out on their own Russian text corpus, which contained 71,000 sentences. They used only n-grams and orthographic features of tokens without utilizing any knowledge-based features. They achieved 89.89% of F-score on three named entity types: names (93.15%), geographical objects (92.7%), and organizations (83.83%).

In [19] the experiments utilized the open Russian text collection "Persons-600"[1] for the person name recognition task. The CRF-based classifier employed such features as token features, context features, and the features based on knowledge about persons (roles, professions, posts, and other). They achieved 88.32% of F-score on person names.

In [10] the experiments were carried out on the Russian text collection, which contained 97 documents. The authors used two approaches for the named entity recognition: knowledge-based and CRF-based approach. In the machine learning framework they utilized such features as the token features and the knowledge features based on word clustering (LDA topics [17], Brown clusters [3], Clark clusters [6]). They achieved 75.05% of F-score on two named entity types: persons (84.84%) and organizations (71.31%).

In 2016 the FactRuEval competition for the Russian language was organized. The FactRuEvaltasks included recognition of names in Russian news texts, recognition of specific attributes of names (family name, first name, etc), and extraction of several types of facts [2].

So far, named entity recognition in tweets did not have studied for Russian. Also, the dependence of NER performance on the language of specific Twitter user communities has not been studied before.

3 Text collections

3.1 News Text Collection

We study the transfer of CRF-based NER classifier trained on newswire data to the tweet collection. For training our system, we chose open Russian text collection "Persons-1000", which contains 1000 news documents labeled with three types of named entities: persons, organizations and locations[2]. The labeling rules are detailed in [16]. The counts of each named entity type in the collection are listed in Table 1.

Table 1 The quantitative characteristics of the labeled named entities in text collections

Type	News collection	Twitter collection
PER	10623	1546
ORG	8541	1144
LOC	7244	2836
OVERALL	26408	5526

3.2 Tweet Text Collection

We are interested in study of the language of Islam-related Twitter users in Russian. To extract tweets from users discussing Islam-related issues, we created a list of 2700 Islam terms. Then we extracted Russian tweets mentioning these terms using Search Twitter API, got users' accounts containing extracted tweets and ordered the accounts in the decreased number of extracted tweets from these accounts. We found that a lot of words from our list practically are not mentioned in tweets, other words (for example, "mosque" or "Muslim") are often used by very different people, not only Muslims.

After studying tweets from extracted accounts we created a very small list of the main Islam words ("Allah", "Quran", "Prophet", in various forms of Russian morphology). We also added the names of several known Islamist organizations to find their possible non-Muslim proponents. Then we repeated the whole procedure of tweet and account extraction, and found that the extracted collections can be considered as an appropriate approximation of messages generated by Islam-related users.

[1]http://ai-center.botik.ru/Airec/index.php?option=com_content&view=article&id=27:persons-600&catid=15&Itemid=40

[2]http://labinform.ru/pub/named_entities/descr_ne.htm

We selected 100 users with the largest number of the extracted tweets, downloaded all their tweets and obtained tweet collection consisting of 300 thousand tweets (further FullTweetCollection). Then we randomly extracted tweets from different users, removed non-Russian or senseless tweets and at last obtained the tweet collection of 4192 tweets (further TestTweetCollection). The created collection contains messages with Quran quotes, religious and political argumentation, news-related messages mainly about Near and Middle East events (Syria, Iraq, Afghanistan etc) and Islamist organizations (Syrian opposition groups, ISIL, etc.) and also other types of messages (for example, advertisements).

The obtained collection was labeled similar to "Persons-1000". To annotate numerous mentions of Allah, we added the Deity type to the annotation scheme, but in the current study we consider the Deity type as a subtype of the Person type.

Analyzing the created collection from the point of view of NER difficulties we found that violations in capitalization mainly include all-caps words for the whole tweet and its fragment. Such capitalization is used for emphasizing important words in the text or words related to Allah as in the following example: За все потери ОН дает нам большую награду" ("For all the losses He gives us a great reward"). Also the tweets mention a lot of Eastern names of persons, organizations ("Фастаким Кама умирт" (Fastakim Kama Umirt group), Джабхатфатхаш-Шам (Jabhat Fateh al-Sham)), or local places difficult for correct recognition.

The fraction of tweets with spelling mistakes, unusual shortenings is relatively low. We suppose that this is because the selected users are well-educated, they are professional writers in some sense, in most cases they are leaders of opinions, whose messages are retweeted by many other people. Therefore it is especially useful to study the specific features of their tweet language.

4 Description of NER Model

In our study, we employ the baseline CRF-classifier that utilizes token features, context features, and lexicon features for NER. Then we consider the ways to improve the baseline model adapting it to the Twitter language. The adaptation techniques include the use of two stage-processing and unsupervised word clustering. Besides, we test the impact of tweet normalization on the NER performance.

4.1 Baseline model

Before named entity recognition with CRF, tweets are processed with a morphological analyzer for determining the part of speech, gender, lemma, grammatical number, case and characteristics of words. This information is used to form features of each word for classifying. In the baseline model we consider three types of features: local token features, context features, and features based on lexicons.

4.1.1 Token features

The token features include:

- Token initial form (lemma);
- Number of symbols in a token;
- Letter case. If a token begins with a capital letter, and other letters are small then the value of this feature is "Big Small". If all letters are capital then the value is "Big Big". If all letters are small then the value is "Small Small". In other cases the value is "Fence";
- Token type. The value of this feature for lexemes is the part of speech, for punctuation marks the value is the type of punctuation;
- Symbol n-grams. The presence of prefixes, suffixes and other n-grams from the predefined sets in a token.

4.1.2 Context-based features

The group of context features includes two feature types. The first type is local context features. It takes into account all mentioned token feature values of nearby words in two-word window to the right and to the left from the current word.

The second type is the bigram context feature. It contains information about the determined named entity type of the previous word. It helps to find named entity borders more precisely. For example, if the person second name is difficult for recognition, the presence of the first name before this word makes the classification easier.

4.1.3 Features based on lexicons

To improve the quality of recognition, we added special lexicons with lists of useful objects. An object can be a word or a phrase. The lexicons had been created before the current work and were not changed during the study.

To calculate the lexicon features, the system matches the text and lexicon entries. If a token is met in a matched lexicon entry then it obtains the lexicon feature value equal to the length of the found entry. The use of the entry length as a feature helps to diminish the affect of lexical ambiguity. For example, in the list of organizations there is "Apple" as the name of a company. But this word does not necessarily mean a company because it has the second sense of a fruit. In the opposite, if we found in the text the phrase "Lomonosov Moscow State University", which is also included in the organization lexicon, the probability of the organization sense is higher than in the first case. The lexicon feature containing the matched entry length helps the system to distinguish these two cases.

The biggest lexicons are listed in Table 2. The overall size of all vocabularies is more than 335 thousand entities. These lexicons were collected from from several sources: phonebooks, Russian Wikipedia, RuThes thesaurus [15].

Table 2 Vocabularysizes

Vocabulary	Size, objects	Clarification
Famous persons	31482	Famous people
First names	2773	First names
Surnames	66108	Surnames
Person roles	9935	Roles, posts
Verbs of informing	1729	Verbs that usually occur with persons
Companies	33380	Organization names
Company types	6774	Oranization types
Media	3909	Media
Geography	8969	Geographical objects
Geographical adjectives	1739	Geographical adjectives
Usual words	58432	Frequent Russian words (nouns, verbs, adjectives)
Equipment	44094	Devices, equipment, tools

4.2 Adaptation of NER Model to Tweets

4.2.1 Unsupervised word clustering

In previous studies it was shown that unsupervised word clustering on the basis of a large text collection improves the NER performance. In our case we compare the impact of word clusters calculated on a large news collection and large tweet collection. For clustering we use the Word2vec package[1]. It represents words with distributional vectors (word embeddings), computed with the neural network. The semantic similarity between two words is calculated as the cosine measure between two corresponding vectors. The package allows automatic clustering of words according to their calculated similarity. We used the c-bow model with vector sizes equal to 300. Thus, each word has an additional feature – the number of a cluster in that it appears. The news collection utilized for clustering contains two million news documents. For tweet-based clustering we use a tweet collection consisting of randomly extracted Russian tweets and including 8.3 million tweets.

4.1.1 Two-stage prediction

We suppose that for adapting a classifier to a text collection it can be useful to take into account the entities already labeled by the classifier and to memorize the named entity type statistics for future use.

On the first stage the classifier extracts named entities. Then the system collects the class statistics determined in the first stage for each word and used it for features of the second stage. After that, new features together with old ones participate in final classification. These statistics can be collected from the current text (the whole text or its part preceding to the word analysis) or from a large text collection (collection statistics). In case of tweet processing, texts are small therefore only the collection statistics can be used. In our experiments this statistics can be obtained from the FullTweetCollection gathered from the selected user accounts or the labeled TestTweetCollection as described in Section 3.2.

For each word, the system finds all mentions of this word in the processed collection and counts frequencies of determined named entity types for this word. Using these frequencies for each entity type, the system creates additional features, which have one of three values: no_one (if the word has not been recognized as a named entity of the chosen type), best (if the word has been as signed to the chosen named entity type more than in 50% of cases), and rare (if the word has been assigned to the chosen named entity type less than in 50% of cases).

For example, if the word “Russia” was met 500 times in a collection, and the classifier assigned it 200 times to organizations and 300 times to locations, then the values of the global statistics feature for the word “Russia” will be as following: PER –no_one, ORG – rare, LOC – best.

4.2 Normalization of Word Capitalization

As we found that in our tweet collection the share of misprints is not very high we did normalization only for word capitalization. The normalization was based on the large news collection described in Section 4.2. For each word in this collection, we counted how many times the word was written in letter case or capital case when it stands not in the beginning of a sentence. The more frequent case was considered as normal for this word.

We considered the normalization in two variants:

- Variant A. All words in a tweet, except the first one, are changed to a normal form of capitalization;
- Variant B. All words in a tweet including the first one are changed to a normal form of capitalization.

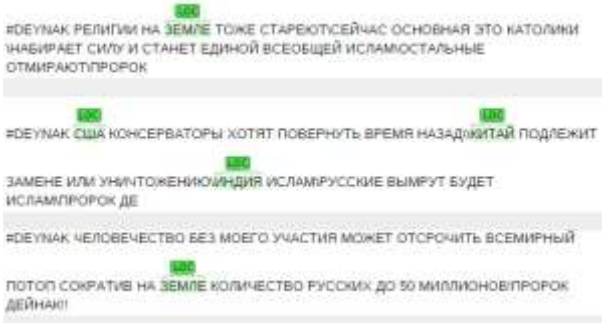


Figure 1 Tweets before normalization

We found that the variant B produces better results and later experimented only with this variant.

Fig. 1 presents several tweets with the manual annotation before normalization. Fig. 2 shows the same tweets after normalization.

[1]https://github.com/dav/word2vec

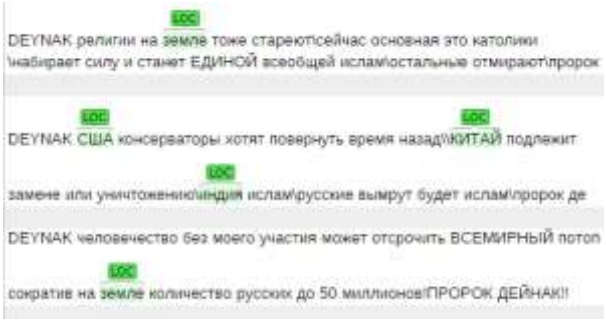


Figure 2 Tweets after normalization

Also the hashtag symbols were removed from a word if this word was found in the news collection to improve its matching with the lexicons.

4 Experiments

In prepocessing we remove mentioned user accounts with "@" and urls in the end of tweets. We consider these data as additional, as meta-information, from which we should not extract names.

We train the described variants of our NER model on the news collection. Table 3 shows results of named entity recognition on the "Persons-1000" collection (cross-validation 3:4). It can be seen that our baseline model is quite good on the news collection and slightly improved after adding clustering features and the two-step approach. In this case the collection statistics is obtained from the same "Persons-1000" collection.

Table 3 News Collection NER Performance

Model	F-measure, %
Baseline	92.49
Baseline+ News clusters	93.48
Baseline+ News clusters + Collection statistics	**93.53**

Then we apply the trained model to the test tweet collection in initial capitalization and normalized capitalization. Table 4 presents the performance of NER models trained on the "Persons-1000" collection for the tweet data. One can see that all models significantly degrade on the tweet collection.

The normalization significantly improves the performance of NER (in contrast to other studies [7]). Word clustering and the collection statistics improve both NER for initial and normalized text collections. Their impact is larger than for the news collection (Table 3). The combination of tweet and news clusters was better than only tweet clusters possibly because of the political and religious character of the gathered collection. In total, the NER performance improves more than 10% on tweet data.

Analyzing mistakes of the best model on the normalized collection we can see still significant share of mistakes because of incorrectly normalized capitalization. We can enumerate the following main subtypes of such problems:

- ambiguous words with different capitalization ("Earth", "Rose"),
- words that should be capitalized in this specific collection. For example, "Paradise" and "Hell" seem to be specific entities in this genre of texts,
- multiword expressions in which each word is usually written in letter case, but together the multiword expression denotes a name and al least the first word should be capitalized. For example, the expression "Московский регион" (Moscow region) is normalized incorrectly because the word "московский" is written in letter case more frequently in the Russian news collection.

Table 4 TweetPerformance

Model	F-measure, TestTweet-Collection	F-measure, Normalized TestTweet-Collection
1) Baseline	64.44%	69.88%
2) Baseline + Collection statistics (TestTweetCollection)	64.99%	70.32%
3) Baseline + Collection statistics (FullTweetCollection)	65.78%	70.44%
4) Baseline + news clusters	66.03%	70.88%
5) Baseline + tweet clusters	66.08%	70.36%
6) Baseline + tweet and news clusters	66.23%	70.89%
7) (2) + tweet and news clusters	**67.27%**	**71.20%**
8) (3) + tweet and news clusters	66.46%	69.73%

5 Conclusion

The paper describes an approach to creating a domain-specific tweet collection written by users frequently discussing Islam-related issues in Russian. We use this collection to study specific features of named entity recognition on Twitter. We found that in contrast to tweets collected randomly, our tweet collection contains relatively small number of spelling errors or strange word shortenings. Specific difficulties of our collection for named entity recognition include a large number of Arabic and other Eastern names (persons, locations, organizations) and frequent use of ALL-CAPS writing for emphasizing main words in messages.

We have studied the transfer of NER model trained on a newswire collection to the created tweet collection and approaches to decrease the degradation of the model because of the transfer. We found that for our specialized text collection, the most improvement was based on normalizing of word capitalization. Two-stage approaches to named entity recognition and word2vec-based clustering were also useful for our task.

In future we plan to improve techniques of tweet normalization and study NER for tweets of followers of the selected users.

References

[1] Antonova, A.Y., Soloviev, A.N.: Conditional Random Field Models for the Processing of Russian. In: Int. Conf. "Dialog 2013", pp. 27- 44. RGGU (2013)

[2] Bocharov, V.V. et al.: "FactRuEval 2016: Evaluation of Named Entity Recognition and Fact Extraction Systems for Russian". In: Dialog Conference. (2016)

[3] Brown, P.F., Della Pietra, V.J., Desouza, P.V., Lai, J.C., Mercer, R.L.: Class-based n-gram Models of Natural Language. Computational Linguistics, 18 (4), pp. 467-479 (1992)

[4] Cherry, C., Guo, H.: The Unreasonable Effectiveness of Word Representations for Twitter Named Entity Recognition. In: NAACL-2015. pp. 735-745 (2015)

[5] Chrupala, G.: Efficient Induction of Probabilistic Word Classes with LDA. In: 5th Int. Joint Conf. on Natural Language Processing, IJCNLP 2011, pp. 363-372. Asian Federation of Natural Language Processing (2011)

[6] Clark, A.: Combining Distributional and Morphological Information for part of Speech Induction. In: 10th Conf. on European Chapter of the Association for Computational Linguistics, EACL, 1, pp. 59-66. ACL (2003)

[7] Derczynski, L., Maynard, D., Rizzo, G., van Erp, M., Gorrell, G., Troncy, R., Bontcheva, K.: Analysis of Named Entity Recognition and Linking for Tweets. Information Processing & Management, 51 (2), pp. 32-49 (2015)

[8] Finin, T., Murnane, W., Karandikar, A, Keller, N., Martineau, J., Dredze, M.: Annotating Named Entities in Twitter Data with Crowdsourcing. In: the NAACL HLT 2010 Workshop on Creating Speech and Language Data with Amazons Mechanical Turk, pp. 80-88 (2010)

[9] Fromreide, H., Hovy, D., Sogaard, A. Crowdsourcing and Annotating NER for Twitter #drift. In LREC-2014, pp. 2544-2547 (2014)

[10] Gareev, R., Tkachenko, M., Solovyev, V., Simanovsky, A., Ivanov, V.: Introducing Baselines for Russian Named Entity Recognition. In: 14th Int. Conf. CICLing 2013, pp. 329-342. Springer (2013)

[11] Hidayatullah, A.F.: Language Tweet Characteristics of Indonesian Citizens. In: Int. Conf. IEEE-2015. pp. 397-401 (2015)

[12] Hovy, D.: Demographic Factors Improve Classification Performance. In: ACL-2015, pp. 752-762 (2015)

[13] Khoroshevsky, V.F.: Ontology Driven Multilingual Information Extraction and Intelligent Analytics. Web Intelligence and Security. pp. 237-262 (2010)

[14] Kuznetsov, I.P., Kozerenko, E.B., Kuznetsov, K.I., Timonina, N.O.: Intelligent System for Entities Extraction (ISEE) from Natural Language Texts. In: Int. Workshop on Conceptual Structures for Extracting Natural Language Semantics-Sense, (9), pp. 17-25 (2009)

[15] Nadeau, D., Sekine, S.: A Survey of Named Entity Recognition and Classification. LingvisticaeInvestigationes, 30 (1), pp. 3-26 (2007)

[16] Paris, C., Thomas, P., Wan, S.: Differences in Language and Style Between Two Social Media Communities. In: the 6th AAAI Int. Conf. on Weblogs and Social Media, ICWSM (2012)

[17] Podobryaev, A.V.: Persons Recognition Using CRF Model. In: 15th All-Russian Scientific Conf. "Digital Libraries: Advanced Methods and Technologies, Digital Collection", RCDL-2013, pp. 255-258. Demidov Yaroslavl State University (2013)

[18] Ratinov, L., Roth, D.: Design Challenges and Misconceptions in Named Entity Recognition. In: 13th Conf. on Computational Natural Language Learning, CoNLL, pp. 147-155. ACL (2009)

[19] Ritter, A., Clark, S., Mausam, Etzioni, O.: Named Entity Recognition in Tweets: An Experimental Study. In: EMNLP, pp.1524-1534 (2011)

[20] Ritter, A, Etzioni, O, Clark, S. et al: Open Domain Event Extraction from Twitter. In: Conf. on Knowledge Discovery and Data Mining, KDD, pp. 1104-1112 (2012)

[21] Cherry, C., Guo, H.: The Unreasonable Effectiveness of Word Representations for Twitter Named Entity Recognition. In: NAACL-2015. pp. 735-745 (2015)

[22] Trofimov, I.V.: Person Name Recognition in News Articles Based on the Persons-1000/1111-F Collections. In: 16th All-Russian Scientific Conf. "Digital Libraries: Advanced Methods and Technologies, Digital Collections", RCDL 2014, pp. 217-221 (2014)

[23] Yang, Y., Eisenstein, J.: Putting Things in Context: Community-specific Embedding Projections for Sentiment Analysis. arXiv preprint arXiv:1511.06052 (2015)

Сравнительный анализ методов автоматической классификации поэтических текстов на основе лексических признаков

© В.Б. Барахнин © О.Ю. Кожемякина © И.С. Пастушков

Институт вычислительных технологий СО РАН,
Новосибирский государственный университет,
Новосибирск, Россия

bar@ict.nsc.ru olgakozhemyakina@mail.ru pas2shkov.ilya@gmail.com

Аннотация. Проанализированы принципы формирования обучающих выборок для алгоритмов определения стилей и жанровых типов. Проведены вычислительные эксперименты с использованием корпуса текстов лицейской лирики А.С. Пушкина по выбору наиболее точного алгоритма классификации поэтических текстов, в том числе с использованием наиболее известных приемов ансамблирования базовых алгоритмов в композиции, таких, как взвешенное голосование, бустинг и стекинг, причем в качестве характеристических признаков стихотворений использовались одиночные слов, биграммы и триграммы. Рассмотренные алгоритмы показали свою работоспособность и могут быть использованы для автоматизации комплексного анализа русских поэтических текстов, существенно облегчая работу эксперта при определении их стилей и жанров путем предоставления соответствующих рекомендаций.

Ключевые слова: автоматический анализ поэтических текстов, определение жанров и стилей, алгоритмы классификации.

Comparative Analysis of Methods of Automated Classification of Poetic Texts Based on Lexical Signs

© V.B. Barakhnin © O.Yu. Kozhemyakina © I.S. Pastushkov

Institute of Computational Technologies of SB RAS,
Novosibirsk, Russia
Novosibirsk State University,
Novosibirsk, Russia

bar@ict.nsc.ru olgakozhemyakina@mail.ru pas2shkov.ilya@gmail.com

Abstract. In this paper we analyze the principles of formation of the training samples for the algorithms of the definition of styles and genre types. The computational experiments with a corpus of texts of Lyceum lyrics of A. S. Pushkin at the choice of the most accurate algorithm of classification of poetic texts were conducted, including the usage of the best-known methods of assembling of the basic algorithms in the composition, such as weighted voting, boosting and stacking, and as a characteristic feature of the poems the single words, bigrams and trigrams were used. The considered algorithms showed their efficiency and can be used to automate the complex analysis of Russian poetic texts, significantly facilitating the work of the expert in determining of their styles and genres by providing the appropriate recommendations.

Keywords: automated analysis of poetic texts, the definition of genres and styles, classification algorithms.

1 Введение

В задачах автоматизированного анализа текстов на естественном языке возникает проблема их классификации по жанрам и стилям, которые являются важными атрибутами, используемыми при определении влияния низших уровней стиха на высшие (см., например, [1]).

Исследования в области автоматизированного определения жанрового типа текстов начаты недавно – в начале 2010-х годов. Так, в работе [2] предложены алгоритмы определения жанров оды,

Труды XIX Международной конференции «Аналитика и управление данными в областях с интенсивным использованием данных» (DAMDID/ RCDL'2017), Москва, Россия, 10–13 октября 2017 года

песни, послания, элегии и эпитафии на материале английских поэтов–сентименталистов XVIII века: поскольку «несмотря на то, что в XVIII–XIX веках жанровые признаки стихотворных текстов постепенно начинают теряться ..., в английской литературе начала XVIII века жанры оды, песни, послания, элегии и эпитафии по соотношению своих формальных признаков еще достаточно хорошо разграничиваются».

В [3] изложен метод классификации текстов (по определенным жанрам и по авторам) на основе анализа статистических закономерностей буквенных распределений, т. е. вероятностей встречаемости букв и буквосочетаний, при этом подчеркнуто, что решение найдено без «вторжения в область литературы, т. е. без анализа синтаксиса, литературных приемов и схем взаимодействий персонажей». Однако в работе [4] сами авторы строят оригинальный контрпример к статистическому методу идентификации, что показывает необходимость использования, по крайней мере, методов морфологического анализа. Что же касается автоматизации определения стилистических характеристик текстов, то нам неизвестны исследования в этой области, по крайней мере, для текстов на русском языке.

В работе [5] нами показано, что метод опорных векторов (support vector machine, SVM) [6] позволил получить хорошие результаты при определении стилистической окраски поэтических текстов и удовлетворительные – при определении жанров.

В настоящей работе мы расширили используемые подходы, в частности, учитывая при построении характеристического вектора используемых в стихотворении лексем количество их вхождений, а также проводя эксперименты с характеристическими векторами биграмм и триграмм. Кроме того, нами был проведен сравнительный анализ целого ряда алгоритмов классификации поэтических текстов, в том числе с использованием наиболее известных приемов ансамблирования, т. е. построения композиций алгоритмов, в которых ошибки отдельных алгоритмов взаимно компенсируются. При ансамблировании рассматриваются алгоритмы, в которых функция, называемая алгоритмическим оператором, устанавливает соответствие между множеством объектов и пространством оценок, а функция, называемая решающим правилом, устанавливает соответствие между пространством оценок и множеством значений целевой функции. Таким образом, рассматриваемые алгоритмы имеют вид суперпозиции алгоритмического оператора и решающего правила. Многие алгоритмы классификации имеют именно такую структуру: сначала вычисляются оценки принадлежности объекта классам, затем решающее правило переводит эти оценки в номер класса. Значением оценки может быть вероятность принадлежности объекта классу, расстояние от объекта до разделяющей поверхности, степень уверенности классификации и т. п.

Таким образом, в статье проведен сравнительный анализ целого ряда методов автоматизированной классификации поэтических текстов, включая наиболее известные приемы ансамблирования базовых алгоритмов в композиции: взвешенное голосование, бустинг и стекинг.

2 Построение обучающей выборки

Наиболее эффективным подходом к автоматизации определения жанровых типов и стилистических характеристик является использование алгоритмов с обучением. Однако формирование обучающей выборки является отнюдь не банальной задачей. Наша попытка использовать в качестве обучающей выборки пушкинскую лирику зрелого периода (1828–1831 гг.) потерпела неудачу уже на раннем этапе работы, поскольку жанровое разнообразие пушкинского творчества этого периода, соотносясь со стилевыми особенностями произведений в особой пушкинской манере, не следует общепринятым законам. На данную черту указывал ране В.А. Грехнев: «Жанры и стиль не противостоят друг другу как враждебные, взаимоотрицающие начала, но между ними всегда существует внутреннее напряжение. Напряжение это возрастает там, где возрастают мощь и размах писательской индивидуальности» [7. С. 234]. Отсюда возникают жанрово-стилистические разновидности и варианты, во «внутреннем напряжении» между стилем и жанром берут начало неканонические жанры, и именно для обучающей выборки это становится критичным, поскольку возникают особенности, не попадающие в систему, следовательно, противоречащие по своей сути материалу для построения жанрово-стилевой системы. Вследствие этого мы решили остановиться на лицейской лирике (1814–1817 гг.), поскольку в ней наблюдаются использование наиболее строгих жанровых форм, стилистическое единство, а также следование правилам грамматик своего времени: «Почти вся лицейская лирика относится к возвышенному стилю, исключение – всего несколько стихотворений. Даже многие сатирические стихи написаны вполне в возвышенном стиле. Можно утверждать, что в ранних стихах Пушкина чувствуется влияние жестких правил «Грамматики» его лицейского учителя Н.Ф. Кошанского» [8. С. 24].

В свою очередь, использование именно лицейской лирики, как материала для создания обучающей выборки, оправдано и стилевым аспектом, поскольку стилевая дифференциация лексем – этап разработки классификатора. Для текстов на русском языке принято восходящее к трудам М.В. Ломоносова [9] деление текстов (прежде всего, художественных) на относящиеся к высокому, среднему и низкому стилям. Исторически каждый из них характеризуется специфическим соотношением использования старославянских (церковнославянских) и собственно русских слов (при этом отдельно рассматривается группа слов,

общих для старославянского и русского языков), долей архаизмов, а также употреблением определенных синтаксических конструкций.

Для реализации поставленной задачи мы идем от практики, делая выборку произведений Пушкина лицейского периода, с 1813 по 1817 гг., как материала, на котором вероятно построение наиболее точной теоретической модели жанрово-стилистических зависимостей, что, несомненно, делает конечный результат анализа наиболее точным и позволяет разработать наиболее адекватный классификатор, относящийся к стилевому аспекту. Так как мы решили ограничиться анализом жанров только малых стихотворных форм, то из анализа исключены поэмы, сказки, переводы, Dubia, и далее делаем список, включающий в себя стихотворения, как соответствующие системе жанров, приведенной в монографии Д.М. Магомедовой [10], так и не входящие в эту систему. В итоге рассмотрения списка произведений, взятого нами для анализа, мы выделяем следующие группы жанров.

Канонические: ода – 4 произведения, элегия – 27 произведений (в том числе одна историческая элегия – «Наполеон на Эльбе»), идиллия – 2 произведения, послание – 55 произведений, баллада – 3 произведения, неканонические, выделенные Д.М. Магомедовой (фрагмент, рассказ в стихах) – их нет.

Также мы добавляем жанры, которых нет в разработанной Д.М. Магомедовой системе канонических–неканонических: эпиграмма – 18 произведений, мадригал – 4 произведения, сонет – 1 произведение, романс – 1 произведение, анекдот – 1 произведение, притча – 2 произведения. Кроме этого, стихотворение «Безверие» (1817) определяется как элегия и философская ода [11]. Но для анализа мы определяем его как философскую оду. Жанровые типы этих произведений легли в основу классификатора (см. табл. 1): по одной оси мы разместили жанровые типы – в порядке возрастания «возвышенности»: ода, элегия, идиллия, послание и т. д., а по другой оси – традиционные стили.

На данном эмпирическом материале просматривается очевидная корреляция между жанровыми и стилистическими характеристиками текстов: ода, элегия и идиллия обычно написаны высоким стилем, в них не используется лексика, соответствующая низкому стилю, а для эпиграмм, напротив, характерно использование элементов лексики низкого стиля. Вообще говоря, стиль текста определяется по наиболее «низким» его лексемам, что особенно характерно для эпиграмм: наличие высокой лексики, употребляемой нередко в ироническом ключе, не должно вводить в заблуждение, ибо употребление одного–двух слов разговорной или откровенно обсценной лексики сразу характеризует авторский замысел. Тем не менее, для жанров, традиционно предполагающих возвышенную форму, прежде всего, мадригала, мы не считаем целесообразным относить принадлежащие к ним стихотворения, в которых с ироническим целями употреблено несколько «сниженных» (но не обсценных!) слов, к сниженному стилю. Следует отметить, что специфика стиля проявляется на лексическом уровне в гораздо большей степени, чем жанр.

В нашей выборке в силу ее специфических задач произведения, написанные в жанре притчи, отнесены: одно («Наездники») – к высокому стилю, второе («Истина») – к среднему, хотя, как известно, притча, будучи жанром, наиболее близким к басне, предполагает возможность написания ее в разных стилях, о чем свидетельствует, в частности, притча Пушкина «Сапожник», которую можно отнести, скорее, к низкому («разговорному») стилю.

Таблица 1 Статистика по жанрово-стилевому соответствию

	Высок.	Средн.	Низк.
Ода	4	-	-
Притча	1	1	-
Мадригал	4	-	-
Послание	-	55	5
Идиллия	-	2	-
Элегия	-	37	-
Романс	-	1	-
Баллада	-	3	-
Эпиграмма	-	-	18
Анекдот	-	-	1

3 О возможности создания словаря стилистически дифференцированных лексем

Прежде, чем приступить к выбору алгоритмов определения стилистических и жанровых характеристик поэтических текстов, необходимо решить вопрос: возможно ли использовать для решения этой задачи априори составленные словари лексем, имеющих ту или иную стилистическую или жанровую окраску?

Большое внимание вопросам стилистической дифференциации слов уделено в монографии О.С. Ахмановой «Очерки по общей и русской лексикологии» [12]. Приведены списки слов «разговорных», со «сниженной» стилевой характеристикой и с «повышенной» стилевой характеристикой. Однако эти списки далеко не полны и носят, скорее, иллюстративный характер, более того, автор признаёт, что «далеко не все из включенных в них слов будут одинаково убедительными (многие, несомненно, покажутся спорными)», и, наконец, стилистическая окраска некоторых лексем менялась со временем, т. е. эта характеристика, взятая из монографии [12], могла быть иной как для языка XIX века, так и для современного. Поэтому для соотнесения слова с тем или иным стилем в той же монографии предложено использовать анализ их структурно-семантической формы. Так, существительные с суффиксом *-к-*а в разнообразных структурно-семантических

вариантах, а также с различными суффиксами со значением «лица» относятся к «разговорной» или «сниженной» лексике; для «разговорной», в отличие от «сниженной», лексики характерно большое число наречий; для «книжной» лексики характерны заимствованные слова, а для «возвышенной» – славянские со сложной структурой, а также архаизмы и т. п.

Однако все эти наблюдения носят весьма частный характер. Так, слова с суффиксом *-к*-а *пытка, речка, шутка* и т. д. встречаются в стихах Пушкина, относящихся отнюдь не к «низкому» или «разговорному» стилю, то же самое относится к словам *бочка, кружка, пушка* и т. д., в которых *–к* является частью корня, но установление этого факта требует нетривиального этимологического анализа, плохо поддающегося автоматизации. Заимствованные слова с течением времени становятся достоянием всех стилей, и это касается не только «древних» заимствований вроде *лошадь* или *собака*, но и новых: *велосипед*, *танк* и т. п. Славянизмы, в том числе со сложной структурой, могли использоваться, в том числе, для придания стихотворению иронического оттенка (например, «Ода его сиятельству графу Д.И. Хвостову» Пушкина и многочисленные сатирические стихи А.К. Толстого).

Ситуация осложняется еще и тем, что нередко «разговорным» или «сниженным» является не все слово в целом, а лишь один из его лексико-семантических вариантов, а также обретением словом той или иной окраски лишь при вхождении в состав фразеологизма.

Таким образом, вхождение в текст отдельных лексем не может служить достаточно надежным критерием отнесения текста к определенному стилистическому типу.

Тем более, четкое выделение жанровой принадлежности отдельных слов представляется совершенно бесперспективной задачей, и нам неизвестны сколько-нибудь удовлетворительные попытки ее разрешения хотя бы на теоретическом уровне.

Именно поэтому нам представляется наиболее целесообразным определять стилистические и жанровые характеристики поэтических текстов на основании вхождения в них совокупности лексем (включая *n*-граммы), определяемых на базе обучающей выборки.

4 Описание численного эксперимента

Для эксперимента использовался описанный выше корпус текстов лицейской лирики Пушкина, состоящий из 121 стихотворения, размеченных экспертом по жанрам и стилям.

При обучении была проведена лемматизация всех уникальных слов, встречающихся в текстах, и создан словарь их исходных форм. Отдельно был составлен словарь имен собственных, которые удалялись из словаря всех слов, поскольку гипотезы, подобные той, что имена из древнегреческого пантеона присущи только высокому стилю, были опровергнуты, в частности, при подготовке данных для экспериментов. Каждый текст кодировался последовательностью цифр, соответствующей количеству вхождений в него слов из словаря: 0 ставился, если слова нет в тексте, 1 – если слово встречается 1 раз, 2 – если 2 и т.д. Помимо лексических признаков, первоначально предполагалось использование стихотворных характеристик (рифма, размер, стопность и т. п.), но это привело к серьёзному ухудшению качества классификации, поэтому было решено от них отказаться.

Также были собраны словари *n*-грамм (*n*=2, 3), которые не содержали имён собственных, причем *n*-граммы были не упорядоченными внутри себя, поскольку в поэзии очень часто встречается обратный порядок слов.

Далее опишем применявшиеся нами приемы ансамблирования, то есть комбинирования алгоритмов, взаимно улучшающего их свойства.

Во-первых, это – два варианта взвешенного голосования с использованием нескольких классификаторов, в случае hard-голосования решение о классификации того или иного объекта принимается на основании заключения большинства используемых классификаторов, в случае soft-голосования результат определяется, исходя из аргумента максимизации вероятности отнесения классифицируемого объекта к некоторому классу.

Во-вторых, это – бустинг, идея которого состоит состоит в жадном выборе очередного алгоритма для добавления в композицию так, чтобы он лучшим образом компенсировал имеющиеся на этом шаге ошибки. Две основные эвристики бустинга – это фиксация $a_1 b_1(x), \dots, a_{t-1} b_{t-1}(x)$ при добавлении $a_t b_t(x)$, где $a_t = \ln \frac{1-p_t}{p_t}, t = 1, \dots, T$, p_t – частота ошибки базового алгоритма b_t, и гладкая аппроксимация пороговой функции потерь.

Нами были применены наиболее известные примеры бустинга – AdaBoost [13], использующий экспоненциальную аппроксимацию функции потерь, и градиентный бустинг (Gradient boosting) [14]. Среди прочих нами был применён метод опорных векторов (Support Vectors Machine, SVM) [6], усиленный AdaBoost.

Наконец, в-третьих, это – стекинг [15], который основан на применении базовых классификаторов для получения предсказаний (метапризнаков) и использовании их как признаков низшего ранга для некоторого «обобщающего» алгоритма (мета-алгоритма). Иными словами, основной идеей стекинга является преобразование исходного пространства признаков задачи в новое пространство, точками которого являются предсказания базовых алгоритмов. В данном исследовании в качестве мета-алгоритма была взята логистическая регрессия над SVM, градиентным бустингом, многослойным персептроном и голосованиями.

Отметим, что в процессе решения рассматриваемой задачи нам пришлось столкнуться

с проблемой миноритарных классов, которые ясно обозначены в таблице 1. Для решения этой проблемы были применены случайное дублирование элементов миноритарных классов, а также стратегия SMOTE [16], которая основана на идее генерации некоторого количества искусственных примеров, которые были бы «похожи» на имеющиеся в миноритарном классе, но при этом не дублировали их. Для создания новой записи вычисляют разность $d = X_b - X_a$, где X_a, X_b – векторы признаков «соседних» примеров a и b из миноритарного класса, которые находят, используя алгоритм ближайшего соседа [17]. В данном случае необходимо и достаточно для примера b получить набор из k соседей, из которого в дальнейшем будет выбрана запись b. Остальные шаги алгоритма ближайшего соседа не требуются. Далее из d путем умножения каждого его элемента на случайное число в интервале (0, 1) получают $\hat{d}$. Вектор признаков нового примера вычисляется путем сложения X_a и $\hat{d}$. Алгоритм SMOTE позволяет задавать количество записей, которое необходимо искусственно сгенерировать. Степень сходства примеров a и b можно регулировать путем изменения значения k (числа ближайших соседей).

Программное приложение для классификации поэтических текстов реализовано на языке Python с использованием библиотек sklearn (реализация алгоритмов, их композиций и кросс-валидации), imblearn (реализация SMOTE), xgboost (наиболее эффективная реализация градиентного бустинга) и pymorphy2 [18] для приведения слов к нормализованному виду, а также для отсечения имен собственных.

В таблицах 2–7 приведены результаты работы классификаторов и их композиций, полученные при трехэтапной кросс-валидации (трехкратное разбиение корпуса на обучающее и тестовое множества, каждый раз классификатор обучался на обучаемом и оценивался на тестовом множестве). Из таблицы результатов был исключен рекомендуемый при работе со SMOTE метод ближайших соседей, так как он показывал очень низкую точность.

Таблица 2 Лексические признаки + SMOTE для определения стиля

Классификатор	**Средн.**	**Max**	**Min**
SVM AdaBoost	0.88	0.91	0.84
XGBoost	0.83	0.9	0.81
Многосл. перс.	0.85	0.95	0.67
Голосование, hard	0.94	0.95	0.92
Голосование, soft	0.94	0.95	0.92
Стекинг	0.94	0.97	0.92

Таблица 3 Лексические признаки + случайное дублирование миноритарных классов для определения жанра

Классификатор	**Средн.**	**Max**	**Min**
SVM AdaBoost	0.88	0.89	0.86
XGBoost	0.90	0.92	0.89
Многосл. перс.	0.93	0.95	0.91
Голосование, hard	0.92	0.95	0.88
Голосование, soft	0.92	0.96	0.88
Стекинг	0.90	0.93	0.87

Таблица 4 Биграммы + SMOTE для определения стиля

Классификатор	**Средн.**	**Max**	**Min**
SVM AdaBoost	0.95	0.98	0.92
XGBoost	0.92	0.97	0.88
Многосл. перс.	0.96	0.98	0.93
Голосование, hard	0.95	0.98	0.91
Голосование, soft	0.94	0.97	0.88
Стекинг	0.95	0.98	0.90

Таблица 5 Биграммы + случайное дублирование миноритарных классов для определения жанра

Классификатор	**Средн.**	**Max**	**Min**
SVM AdaBoost	0.94	0.96	0.90
XGBoost	0.97	1.00	0.93
Многосл. перс.	0.97	0.99	0.94
Голосование, hard	0.94	1.00	0.88
Голосование, soft	0.93	1.00	0.88
Стекинг	0.96	1.00	0.89

Таблица 6 Триграммы + SMOTE для определения стиля

Классификатор	**Средн.**	**Max**	**Min**
SVM AdaBoost	0.83	0.98	0.88
XGBoost	0.90	0.94	0.87
Многосл. перс.	0.95	0.99	0.93
Голосование, hard	0.93	0.98	0.89
Голосование, soft	0.91	0.98	0.88
Стекинг	0.94	0.99	0.89

Таблица 7 Триграммы + случайное дублирование миноритарных классов для определения жанра

Классификатор	**Средн.**	**Max**	**Min**
SVM AdaBoost	0.95	1.00	0.86
XGBoost	0.94	1.00	0.84
Многосл. перс.	0.97	0.99	0.95
Голосование, hard	0.96	1.00	0.91
Голосование, soft	0.96	1.00	0.91
Стекинг	0.96	1.00	0.88

Из полученных данных можно сделать следующие выводы:

- стекинг не всегда даёт наилучшее (т. е. наиболее соответствующее экспертной оценке) решение (табл. 3);
- при увеличении контекста признаков (от одного слова к би- и триграммам) XGBoost становится более точным, чем многослойный персептрон;

- увеличение ширины контекста приводит к улучшению качества, но только до определённого момента (использование тетраграмм дало заметное ухудшение результатов). Отметим, что применение популярной концепции word2vec [19] дало очень слабый результат (0.83–0.85), и при этом время подсчёта увеличилось в несколько раз;
- на основе лексических признаков или *n*-грамм можно получить хороший результат даже с помощью простых классификаторов;
- исходя из критерия максимизации минимальной точности, следует использовать многослойный персептрон, а в качестве лексических характеристик стихотворений – триграммы.

5 Заключение

В работе проанализированы принципы формирования обучающих выборок для алгоритмов определения стилей и жанровых типов. Проведены вычислительные эксперименты с использованием корпуса текстов лицейской лирики А.С. Пушкина по выбору наиболее точного алгоритма классификации поэтических текстов, в том числе с использованием наиболее известных приемов ансамблирования базовых алгоритмов в композиции, таких, как взвешенное голосование, бустинг и стекинг, причем в качестве характеристических признаков стихотворений использовались одиночные слов, биграммы и триграммы. Рассмотренные алгоритмы показали свою работоспособность (при этом, исходя из критерия максимизации минимальной точности, следует использовать многослойный персептрон, а в качестве лексических характеристик стихотворений – триграммы) и могут быть использованы для автоматизации комплексного анализа русских поэтических текстов, существенно облегчая работу эксперта при определении их стилей и жанров путем предоставления соответствующих рекомендаций.

Поддержка

Работа выполнена при частичной поддержке Президиума РАН (проект 2016-PRAS-0015) и Президентской программы «Ведущие научные школы РФ» (грант 7214.2016.9).

Литература

[1] Барахнин, В.Б., Кожемякина, О.Ю. Об автоматизации комплексного анализа русского поэтического текста. CEUR Workshop Proceedings, 934, сс. 167-171 (2012)

[2] Лесцова, М.А.: Определение ядра и периферии жанров оды, песни, послания, элегии и эпитафии на материале английских поэтов-сентименталистов XVIII века. Вестник Челябинского государственного пед. университета, 4, сс. 196-205 (2014)

[3] Орлов, Ю.Н., Осминин, К.П.: Определение жанра и автора литературного произведения статистическими методами. Прикладная информатика, 26 (2), сс. 95-108 (2010)

[4] Орлов, Ю.Н., Осминин, К.П.: Методы статистического анализа литературных текстов. Эдиториал УРСС, Москва (2012)

[5] Barakhnin, V., Kozhemyakina, O., Pastushkov, I.: Automated Determination of the Type of Genre and Stylistic Coloring of Russian Texts. ITM Web of Conferences 10, 02001, 4 p. (2017). doi: 10.1051/itmconf/20171002001

[6] Vapnik,V.N.: The Nature of Statistical Learning Theory. Springer-Verlag (1995)

[7] Грехнев, В.А.: Лирика Пушкина. О поэтике жанров. Горький: Волго-Вятское книжное издательство (1985)

[8] Барахнин, В.Б., Кожемякина, О.Ю.: К проблеме аутентичности фонетического анализа в связи с возможными особенностями авторской орфографии (на примере чередования окончаний -ой/-ый в лирике А.С. Пушкина). Вестник Томского государственного университета. Филология, 13 (2), сс. 5-28 (2016)

[9] Ломоносов, М.В.: Предисловие о пользе книг церковных в российском языке. Ломоносов, М. В. Полн. собр. соч. 7, сс. 585-592. М.–Л.: Изд-во АН СССР (1952)

[10] Магомедова, Д.М.: Филологический анализ лирического стихотворения. М.: Издательский центр «Академия» (2004)

[11] Свободина, С.Ф.: К вопросу о философской направленности и жанровых особенностях стихотворения А.С. Пушкина «Безверие». Пушкинский музеум: альманах, 6, сс. 261-270. Всероссийский музей А.С. Пушкина, Санкт-Петербург (2014)

[12] Ахманова, О.С.: Очерки по общей и русской лексикологии. М.: Учпедгиз (1957)

[13] Freund, Y., Schapire, R.E.: A Short Introduction to Boosting. J. of Japanese Society for Artificial Intelligence, 14 (5), pp. 771-780 (1999)

[14] Friedman, J.H.: Stochastic Gradient Boosting. Computational Statistics and Data Analysis, 38 (4), pp. 367-378 (2002)

[15] Wolpert, D.H.: Stacked Generalization. Neural Networks, 5 (2), pp. 241-259 (1992)

[16] Chawla, N.V.: Data Mining for Imbalanced Datasets: An Overview. Data Mining and Knowledge Discovery Handbook, pp. 875-886. Springer-Verlag (2010)

[17] Cover, T.M., Hart, P.E.: Nearest Neighbor Pattern Classification. IEEE Transactions on Information Theory, 13, pp. 21-27 (1967)

[18] Korobov, M.: Morphological Analyzer and Generator for Russian and Ukrainian Languages. Analysis of Images, Social Networks and Texts. Communications in Computer and Information Science, 542, pp. 320-332 (2015)

[19] Mikolov, T., Kai, Chen, Corrado, G., Dean, J.: Efficient Estimation of Word Representations in Vector Space. Computation and Language, 12 p., (2013). https://arxiv.org/pdf/1301.3781.pdf

Онтологические модели и применения 2

Ontological models and applications 2

Применение паттернов онтологического проектирования при разработке онтологий научных предметных областей

© Ю.А. Загорулько[1,2] © О.И. Боровикова[1] © Г.Б. Загорулько[1,2]

[1] Институт систем информатики имени А.П. Ершова СО РАН,
[2] Новосибирский государственный университет,
Новосибирск, Россия

zagor@iis.nsk.su olesya@iis.nsk.su gal@iis.nsk.su

Аннотация. Обсуждены вопросы применения паттернов онтологического проектирования при разработке онтологий научных предметных областей. Такие паттерны предназначены для описания решений типовых проблем, возникающих при разработке онтологий. Они создаются для того, чтобы облегчить процесс построения онтологий и помочь разработчикам избежать некоторых, часто повторяющихся ошибок моделирования. Представлены паттерны онтологического проектирования, сложившиеся в результате решения проблем, с которыми авторы столкнулись при разработке онтологий для таких научных предметных областей, как археология, компьютерная лингвистика, системные исследования в энергетике, активная сейсмология и др.

Ключевые слова: научная предметная область, онтология, паттерны онтологического проектирования, методология разработки онтологий.

Application of Ontology Design Patterns in the Development of the Ontologies of Scientific Subject Domains

© Yu.A. Zagorulko[1,2] © O.I. Borovikova[1] © G.B. Zagorulko[1,2]

[1] A.P. Ershov Institute of Informatics Systems,
[2] Novosibirsk State University,
Novosibirsk, Russia

zagor@iis.nsk.su olesya@iis.nsk.su gal@iis.nsk.su

Abstract. The paper discusses the application of ontology design patterns for the development of ontologies of scientific subject domains. Such patterns are designed to describe the solutions of typical problems arising in ontology development. They are created to facilitate the process of building ontologies and to help the developers avoid some frequent errors occurring in ontology modeling. The paper presents the ontology design patterns resulting from solving the problems that the authors have encountered in the development of ontologies for such scientific subject domains as archaeology, computational linguistics, system studies in power engineering, active seismology, etc.

Keywords: scientific subject domain, ontology, ontology design patterns, methodology of ontology development.

1 Введение

В настоящее время наиболее популярным и эффективным средством концептуализации и формализации научных предметных областей являются онтологии [12]. Они повсеместно используются для представления и фиксации общего разделяемого всеми экспертами (или группой экспертов) знания о таких областях. Формализация семантики предметной области в виде онтологии служит не только целям компактного и непротиворечивого ее описания, она также формирует понятийный базис для представления всей совокупности знаний о ней. Например, в системе информационной поддержки научной деятельности (СПНД) [19] в терминах онтологии может быть описана семантика используемых в ней данных и информационных ресурсов, а в экспертной системе или системе поддержки принятия решений – экспертные правила, прецеденты и другие компоненты базы знаний [27].

СПНД должна обеспечивать пользователю

Труды XIX Международной конференции «Аналитика и управление данными в областях с интенсивным использованием данных» (DAMDID/ RCDL'2017), Москва, Россия, 10–13 октября 2017 года

представление всей необходимой ему информации об интересующей его области знаний, о ее составляющих (разделы/подразделы науки, объекты, методы и техники исследования и т. п.), а также о субъектах (участниках) научной деятельности (персоналиях, группах, сообществах и других организациях, включенных в процесс исследования). В СПНД онтология задает формальное описание области знаний, на основе которого систематизируется такая информация, выполняется интеграция в единое информационное пространство релевантных информационных ресурсов и документов. На основе онтологии также строится пользовательский интерфейс, обеспечивающий содержательный доступ к знаниям и данным, интегрированным в информационное пространство системы. В частности, в таком интерфейсе пользователь может использовать онтологию в качестве «проводника» для навигации по этому пространству, а также формулировать поисковые запросы, основными элементами которых являются понятия и отношения онтологии.

В настоящее время онтологии широко используются для концептуального моделирования предметных областей с интенсивным использованием данных [31]. Развитие и применение инфраструктур поддержки научных исследований, базирующихся на концептуальных спецификациях таких областей, позволяют избежать зависимости программ от структуры источников данных, обеспечить интероперабельность различных методов обработки данных при совместной работе, повысить надёжность получаемых результатов за счёт использования формальных непротиворечивых спецификаций.

Разработка онтологий научных предметных областей является довольно сложным и трудоемким процессом. Для его упрощения и облегчения предложены различные методы и подходы к разработке онтологий. На протяжении последних десяти лет интенсивно развивается подход, базирующийся на применении паттернов онтологического проектирования (Ontology Design Patterns или ODP) [6, 14].

ODP представляют собой документально зафиксированные описания проверенных на практике решений проблем онтологического моделирования. В настоящее время создано и развивается несколько каталогов паттернов [13, 29]. Следует заметить, что такие каталоги, как правило, ориентированы либо на какую-то предметную область, либо группу разработчиков, поэтому не обладают полнотой и универсальностью.

В статье обсуждаются паттерны онтологического проектирования, сложившиеся в результате решения проблем, с которыми авторы столкнулись в процессе разработки онтологий для различных научных предметных областей [20, 21, 23–25, 32]. Описание проблем и паттернов дано в контексте методологии разработки онтологий для тематических интеллектуальных научных интернет-ресурсов [26], предназначенных для информационной и аналитической поддержки научной деятельности в заданных областях знаний.

2 Обзор паттернов онтологического проектирования

Паттерны онтологического проектирования имеют в качестве своих прародителей паттерны проектирования, широко используемые в разработке программного обеспечения. В этой области деятельности под паттерном проектирования (design pattern) понимается описание хорошо проверенной, обобщенной схемы решения некоторой часто повторяющейся проблемы разработки, которая возникает в некотором контексте. Паттерны вошли в повседневную практику объектно-ориентированного проектирования. С их помощью решаются конкретные задачи проектирования, в результате чего объектно-ориентированный дизайн становится более гибким, элегантным, и повторно используемым [22].

По аналогии с паттернами проектирования, паттерны онтологического проектирования предназначены для описания решений типовых проблем, возникающих при разработке онтологий. Паттерны создаются для того, чтобы облегчить процесс построения онтологий и помочь разработчикам избежать некоторых часто повторяющихся ошибок онтологического моделирования [8]. В таком качестве ODP были впервые независимо друг от друга введены Aldo Gangemi [6] и Eva Blomqvist с Kurt Sandkuhl [4].

Основной каталог паттернов онтологического проектирования представлен на портале Ассоциации ODPA (Association for Ontology Design & Patterns) [29], созданного в рамках проекта NeOn [30]. В рамках этого проекта была предложена представленная ниже типология паттернов [15].

В зависимости от проблем, для решения которых предназначены паттерны онтологического проектирования, различают структурные паттерны (Structural ODPs), паттерны соответствия (Correspondence ODPs), паттерны содержания (Content ODPs), паттерны логического вывода (Reasoning ODPs), паттерны представления (Presentation ODPs) и лексико-синтаксические паттерны (Lexico-Syntactic ODPs) [7].

Структурные паттерны либо фиксируют способы решения проблем, вызванных ограничениями выразительных возможностей языков описания онтологий, либо задают общую структуру и вид онтологии. Паттерны первого типа называются логическими паттернами (Logical ODP). К ним относится, например, паттерн многоместного отношения, решающий проблему отсутствия во многих языках описания онтологий отношений такого вида. Паттерны второго типа – архитектурные паттерны (Architectural ODP) – содержат предложения по организации онтологии в целом, включая, например, такие структуры, как таксономия и модульная архитектура.

Заметим, что структурные паттерны являются

предметно-независимыми, на их основе могут строиться фрагменты онтологии, входящие в паттерны содержания.

Паттерны логического вывода строятся на основе структурных логических паттернов и предназначены для получения определенных результатов с помощью машины логического вывода. Такие паттерны обеспечивают не только вывод неявно заданных в онтологии знаний (паттерны классификации, категоризации, наследования и др.), но и проверку онтологии на непротиворечивость и полноту, выполнение запросов к онтологии, ее оценку и нормализацию (устранение анонимности классов и экземпляров, явное представление (reification) иерархии классов, нормализацию имен и т. п.) [18].

Паттерны содержания задают способы представления типовых фрагментов онтологий, на основе которых могут строиться онтологии целого класса предметных областей.

Паттерны представления определяют рекомендации по именованию, аннотированию и графическому представлению элементов онтологии, применение которых должно повысить понимаемость онтологии, а также удобство и простоту ее использования.

Паттерны соответствия требуются для выполнения реинжиниринга (трансформации) и выравнивания (отображения) онтологий. Первая группа паттернов применяется, когда необходимо построить новую онтологию (при этом исходная модель не обязательно является онтологической). Вторая группа используется для установления соответствий между понятиями и индивидами двух онтологий, чтобы обеспечить интероперабельность без изменения существующих моделей.

Лексико-синтаксические паттерны применяются для облегчения построения (пополнения) онтологий на основе текстов на естественном языке. Они задают отображения языковых структур в онтологические структуры.

Следует заметить, что на данный момент не существует единого стандарта для описания паттернов [11], но чаще всего они описываются в формате, предложенном на портале ассоциации ODPA [29]. Схема описания паттерна включает его графическое представление, текстовое описание, набор сценариев, примеров и ссылки на другие паттерны, в которых он используется, а также общую информацию о названии паттерна, его авторе и области применения. Согласно методологии eXtreme Design [3] каждый паттерн содержания снабжается также набором квалификационных вопросов, определяющих его содержание.

3 Паттерны методологии построения онтологий для тематических ИНИР

В этом разделе рассмотрены структурные и содержательные паттерны, которые применяются в методологии построения онтологий для тематических интеллектуальных научных интернет-ресурсов (ИНИР) [26], в разработке которой принимали участие авторы статьи. Данная методология использует средства технологии Semantic Web [9]. В частности, онтологии в рамках этой методологии разрабатываются на языке OWL [1] с использованием редактора Protégé. Эти средства помогают решить многие проблемы онтологического инжиниринга, включая проверку корректности и повторное использование онтологий, но их применение, в свою очередь, создает новые проблемы.

3.1 Структурные паттерны

Необходимость в использовании структурных логических паттернов в методологии построения онтологий для тематических ИНИР была вызвана проблемой отсутствия в языке OWL выразительных средств для представления сложных сущностей и конструкций, актуальных при построении онтологии тематических ИНИР, в частности, областей допустимых значений, многоместных и атрибутированных отношений (бинарных отношений с атрибутами).

Первым рассмотрим паттерн представления области допустимых значений свойств (см. Рис. 1), введение которого было вызвано проблемой отсутствия в языке OWL специальных средств для задания таких областей, которые в реляционной модели данных называются доменами и характеризуются названием и множеством элементарных значений. Домены удобно использовать при описании возможных значений свойств (атрибутов) класса, когда весь набор таких значений известен заранее. Использование доменов не только позволит контролировать ввод информации, но и может повысить удобство этой операции, за счет обеспечения пользователям возможности выбора значений свойств из заданного списка значений.

Решением указанной проблемы является задание домена перечислимым классом, который является наследником специально введенного служебного класса *Домен*. Конкретный домен не имеет наследников и состоит из конечного набора различных индивидов (объектов или экземпляров класса), определяющих возможные значения заданного свойства (*ObjectProperty*) для объектов рассматриваемого класса (см. Рис. 1).

Примерами таких доменов являются “Географический тип”, “Должность”, “Тип организации”, “Тип публикации”, которые включают соответственно виды населенных пунктов, виды должностей, типы организаций и публикаций (описание домена “Географический тип” представлено в нижней части Рис. 1).

Заметим, что на приведенных в статье рисунках паттернов классы обозначаются в виде эллипсов, а их индивиды и атрибуты – в виде прямоугольников. Связь типа *ObjectProperty* показывается сплошной прямой линией, а связь типа *DataProperty* – прерывистой. При этом обязательные классы и атрибуты (индивиды) представляются фигурами, обведенными жирной линией.

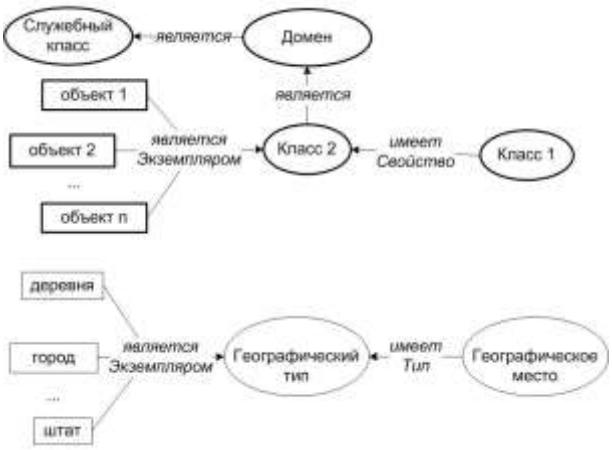


Рисунок 1 Структурный паттерн представления области допустимых значений и пример его использования

Другой часто возникающей проблемой при разработке онтологии является потребность в представлении атрибутированных отношений между объектами. Для этих целей, как правило, используются обычные бинарные отношения, снабженные атрибутами, специализирующими связь между аргументами отношения [5]. Так как в языке OWL нет возможности задания атрибутов для отношений, был предложен структурный паттерн, предусматривающий введение служебного класса *Отношение с атрибутами* (см. Рис. 2).

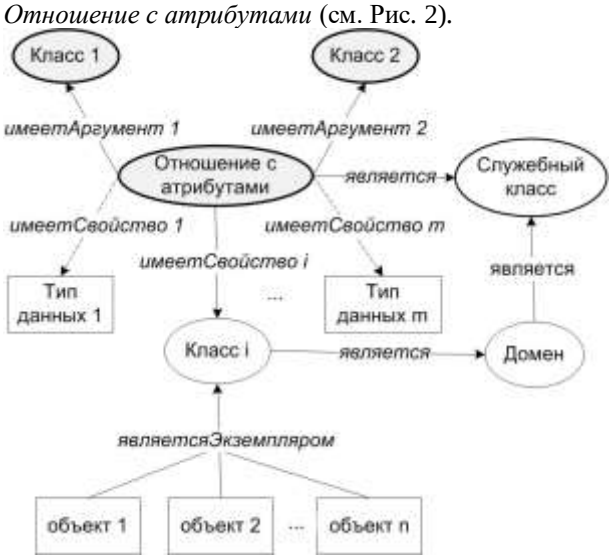


Рисунок 2 Структурный паттерн бинарного атрибутированного отношения

Для представления конкретного типа отношения вводится новый класс – его наследник. Экземпляр этого класса связывается с каждым аргументом и атрибутом атрибутированного отношения. При этом нужно учитывать необходимость задания ограничений обязательности и единственности для аргументов, в то время как ограничения на число атрибутов (свойств) не задаются.

Заметим, что данный паттерн, в отличие от введенного в [5] паттерна *Qualified Relation*, позволяет явно указать порядок аргументов атрибутированного отношения, сохраняя информацию об его ориентированности, что важно для представления пользователю полной информации о характере связи между объектами.

На Рис. 3 представлен пример использования данного паттерна для задания отношения, описывающего участие персоны (класс *Персона*) в некоторой деятельности (класс *Деятельность*). Паттерн позволяет задать даты начала и окончания участия персоны в некоторой деятельности, а также его роль в ней.

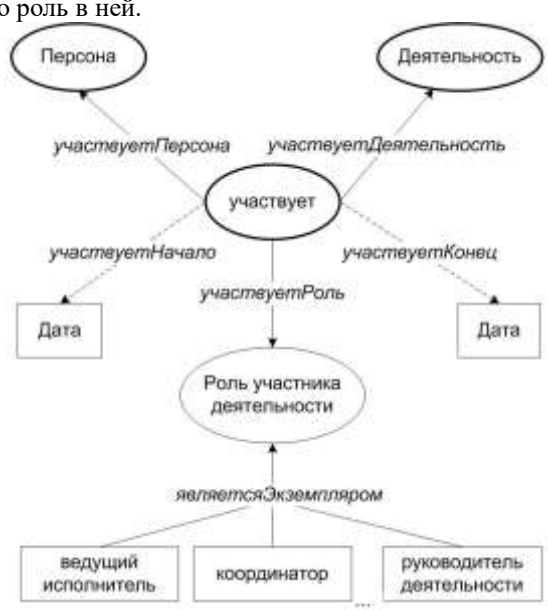


Рисунок 3 Паттерн бинарного атрибутированного отношения *участвует*

Подобным образом можно построить паттерн для многоместного отношения, для классов-аргументов которого, кроме свойства обязательности и единственности, указывается также порядок аргументов (см. Рис. 4).

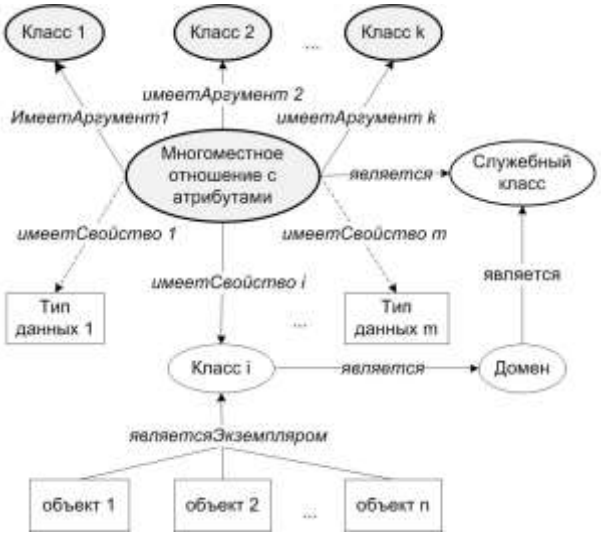


Рисунок 4 Структурный паттерн многоместного отношения с атрибутами

Заметим, что в отличие от структурного паттерна многоместного отношения (*N-Ary Relation Pattern*),

представленного в [10], и содержательного паттерна “Ситуация” (*Situation*), представленного в [17], на основе которого предлагается описывать многоместные отношения, в паттерне, предложенном в данной работе, помимо аргументов отношения и их порядка также можно задавать его свойства (*ObjectProperty*) и атрибуты (*DataProperty*). Это в значительной степени повышает изобразительные возможности данного паттерна.

Структурные паттерны являются предметно-независимыми, на их основе могут задаваться элементы онтологии для паттернов содержания.

3.2 Паттерны содержания

Как было сказано выше, паттерны содержания задают способы представления типовых фрагментов онтологий, на основе которых могут строиться онтологии моделируемых предметных областей. Фактически предлагаемые паттерны содержания являются фрагментами базовых онтологий, предоставляемых упомянутой выше методологией построения онтологий для тематических ИНИР, которые после конкретизации (специализации) содержащихся в них понятий и расширения новыми понятиями становятся составными частями онтологий конкретных предметных областей.

Онтология предметной области ИНИР строится на основе следующих базовых онтологий: онтологии научного знания и научной деятельности, базовой онтологии задач и методов, а также базовой онтологии научных информационных ресурсов [26].

Онтология научного знания содержит классы, задающие структуры для описания понятий, входящих в любую научную область знаний. Такими понятиями являются *Раздел науки, Объект исследования, Предмет исследования, Метод исследования, Научный результат* и др. Эта онтология также включает отношения, связывающие между собой объекты указанных выше классов. Использовав эти классы, можно выделить и описать разделы и подразделы, значимые для моделируемой области знаний, задать типизацию методов и объектов исследования, описать результаты научной деятельности.

Онтология научной деятельности базируется на онтологии, предложенной в [2] для описания научно-исследовательских проектов и расширенной для применения к более широкому классу задач. Эта онтология включает классы понятий, относящиеся к организации научной и исследовательской деятельности, такие, как *Персона* (*Исследователь*), *Организация, Событие, Деятельность, Проект, Публикация* и др. В эту онтологию входят также отношения, позволяющие связывать понятия данной онтологии не только между собой, но и с понятиями онтологии научного знания.

Базовая онтология научных информационных ресурсов в качестве основного класса включает класс *Информационный ресурс*, так как данное понятие играет важную роль во всех научных областях. Набор атрибутов и связей этого класса основан на стандарте Dublin core [16]. Его атрибутами являются: название ресурса, язык ресурса, тематика ресурса, тип ресурса, дата создания ресурса и др. Для представления информации об источниках ресурса и его создателях, а также связанных с ним событиях, организациях, персонах, публикациях и других сущностях вводятся специальные отношения, связывающие класс *Информационный ресурс* с классами других базовых онтологий.

Базовая онтология задач и методов включает такие классы, как *Задача, Метод решения* и *Web-сервис*. С помощью понятий и отношений данной онтологии могут быть описаны задачи, для решения которых предназначен ИНИР, методы их решения и реализующие их веб-сервисы.

Для описания научных предметных областей требуется уметь единообразно представлять используемые в них понятия и их свойства. Для этого были разработаны паттерны для представления основных понятий и отношений базовых онтологий. Покажем, как выглядят паттерны онтологии научного знания.

Паттерн, представленный на Рис. 5, предназначен для описания методов исследования, используемых в научной деятельности.

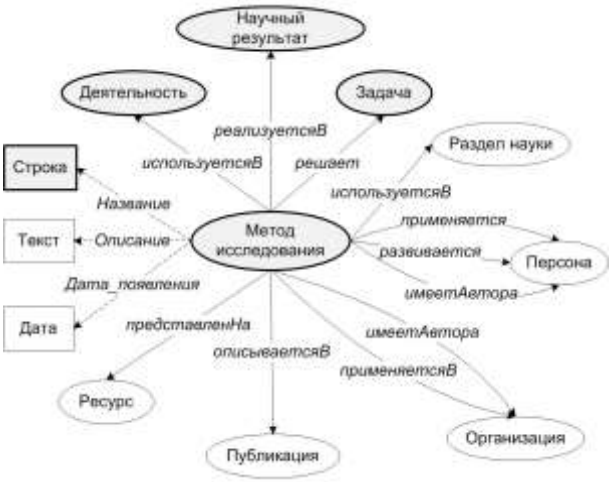


Рисунок 5 Паттерн для описания метода исследования

Элементы описания паттерна метода исследования представлены такими обязательными классами онтологии, как *Деятельность*, *Научный Результат*, *Задача*, и соответствующими отношениями *используетсяВ*, *реализуетсяВ*, *решает*.

Приведем набор квалификационных вопросов, представляющих содержание этого паттерна:

Каково название метода исследования?
В какой деятельности используется метод?
Какие задачи решаются с помощью метода?
В каких разделах науки используется метод?
В каких научных результатах метод реализован?
Когда метод впервые появился?
Кто является автором метода?
Какие персоны применяют метод?
Какие персоны развивают метод?
В каких организациях применяется метод?

В каких публикациях метод описывается?

На каких ресурсах представлен метод?

На Рис. 6 представлен пример использования описанного выше паттерна для описания метода недоопределенных вычислений [28], предложенного А.С. Нариньяни в 1986 г. и реализованного в решателе UniCalc.

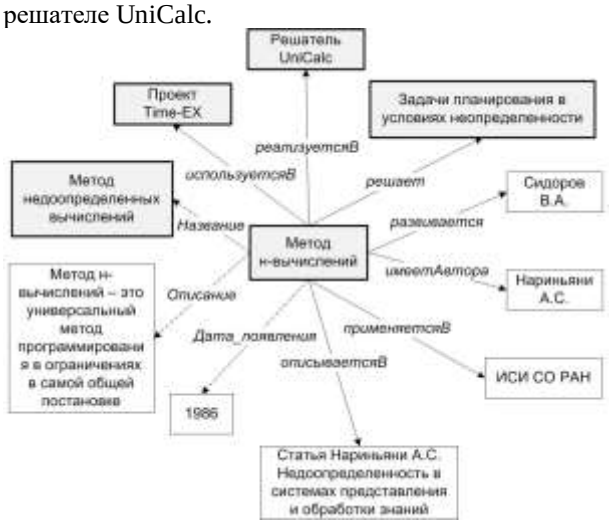


Рисунок 6 Пример использования паттерна метода исследования

На Рис. 7 приведен паттерн для описания объекта исследования, который включает в качестве обязательных следующие классы: *Предмет Исследования*, *Деятельность*, *Раздел науки*. Экземпляры этих классов должны быть связаны с объектом исследования отношениями *имеетАспект*, *исследуетсяВ и изучаетсяВ* соответственно. При этом объект исследования может быть структурным (включать в себя другие объекты исследования).

Паттерн предмета исследования (см. Рис. 8) обязательно должен включать ссылку на объект исследования, аспектом которого он является. Предмет исследования, как и объект исследования, может быть структурным (включать в себя другие предметы исследования).

Рисунок 7 Паттерн для описания объекта исследования

Рисунок 8 Паттерн для описания предмета исследования

В описании научной деятельности важное место занимают научные результаты. Паттерн, предназначенный для описания научного результата, приведен на Рис. 9. В этом паттерне отражено требование, состоящее в том, что при описании научного результата необходимо давать ссылку на деятельность, при выполнении которой он был получен.

Заметим, что в представленных выше паттернах используются не только «центральные» понятия паттернов, но и понятия из смежных паттернов (например, в паттерне описания предмета исследования, кроме понятия *Предмет Исследования*, используются такие понятия, как *Объект исследования, Научный результат, Раздел науки* и др.) Это позволяет давать связанное описание моделируемой области.

Рисунок 9 Паттерн для описания научного результата

3 Заключение

Рассмотрены вопросы применения паттернов онтологического проектирования для разработки

онтологий научных предметных областей. Была описана предложенная Ассоциацией ODPA классификация паттернов и подробно рассмотрены авторские паттерны, которые использовались при разработке онтологий для ряда научных предметных областей.

На основе паттернов обеспечивается согласованное представление всех сущностей онтологии. Использование экспертами и инженерами знаний паттернов онтологического проектирования позволяет сэкономить ресурсы и избежать ошибок при разработке онтологий.

Поддержка

Работа выполнена при частичной финансовой поддержке Российского фонда фундаментальных исследований (грант № 16-07-00569) и Президиума РАН (проект II.2П/IV.39-1 «Информационные, управляющие и интеллектуальные технологии и системы»).

Литература

[1] Antoniou, G., Harmelen, F.: Web Ontology Language: OWL. Handbook on Ontologies, pp. 67-92. Berlin: Springer Verlag (2004)

[2] Benjamins, V. R., Fensel, D.: Community is Knowledge! in (KA)2. Proc. of 11th Banff Knowledge Acquisition for Knowledge-based Systems workshop KAW'98 (Banff, Canada, April 1998), pp. KM.2-1 – KM.2-18. Calgary: SRDG Publications, Department of Computer Science, University of Calgary (1998)

[3] Blomqvist, E., Hammar, K., Presutti, V.: Engineering Ontologies with Patterns: The eXtreme Design Methodology. In: Hitzler, P., Gangemi, A., Janowicz, K., Krisnadhi, A., Presutti, V. (eds) Ontology Engineering with Ontology Design Patterns, Studies on the Semantic Web, 25, pp. 23-50. IOS Press (2016)

[4] Blomqvist, E., Sandkuhl, K.: Patterns in Ontology Engineering: Classification of Ontology Patterns. Proc. of the Seventh Int. Conf. on Enterprise Information Systems ICEIS 2005, Miami, USA, pp. 413-416 (2005)

[5] Dodds L., Davis I.: Linked Data Patterns: A Pattern Catalogue for Modelling, Publishing, and Consuming Linked Data. http://patterns.dataincubator.org/book

[6] Gangemi, A.: Ontology Design Patterns for Semantic Web Content. In: Gil, Y., Motta, E., Benjamins, V.R., Musen, M.A. (eds). The Semantic Web – ISWC 2005. LNCS, 3729, pp. 262-276. Springer: Berlin, Heidelberg (2005)

[7] Gangem, A., Presutti, V.: Ontology Design Patterns. Handbook on Ontologies, pp. 221-243. Springer (2009)

[8] Hammar, K.: Towards an Ontology Design Pattern Quality Model. Linköping Studies in Science and Technology Linköping University, 1606 (2013)

[9] Hitzler, P., Krötzsch, V., Rudolph, S.: Foundations of Semantic Web Technologies. Chapman & Hall/CRC (2009)

[10] Hoekstra, R.: Ontology Representation – Design Patterns and Ontologies that Make Sense. Frontiers of Artificial Intelligence and Applications, 197, pp. 1-236. IOS Press, Amsterdam (2009)

[11] Karima, N., Hammar, K., Hitzler, P.: How to Document Ontology Design Patterns. Proc. of the 7th Workshop on Ontology and Semantic Web Patterns (WOP 2016), Kobe, Japan. IOS Press (2016)

[12] Ontologies: A Handbook of Principles, Concepts and Applications in Information Systems. Sharman, R., Kishore, R., Ramesh, R. (eds). Springer New York, Secaucus, NJ, USA (2006)

[13] Ontology Design Patterns (ODPs) Public Catalog. http://odps.sourceforge.net

[14] Ontology Engineering with Ontology Design Patterns: Foundations and Applications. Studies on the Semantic Web. Hitzler, P., Gangemi, A., Janowicz, K., Krisnadhi, A., Presutti, V. (eds), IOS Press/AKA (2016)

[15] Presutti, V., Gangemi, A., David, S., Aguado de Cea, G., Su´arez–Figueroa, M.C., Montiel-Ponsoda, E., Poveda, M. D2.5.1: A Library of Ontology Design Patterns: Reusable Solutions for Collaborative Design of Networked Ontologies. Technical report, NeOn Project (2007)

[16] Rühle, S., Baker, T., Johnston, P. User Guide. http://wiki.dublincore.org/index.php/User_Guide

[17] Submissions:Situation. http://ontologydesignpatterns.org/wiki/Submissions:Situation

[18] Vrandečić, D., Sure, Y.: How to Design Better Ontology Metrics. Proc. of the Semantic Web: Research and Applications: 4th European Semantic Web Conference, ESWC 2007, Innsbruck, Austria, June 3–7, 2007. LNCS, 4519, pp. 311-325. Springer (2007)

[19] Zagorulko, Yu., Zagorulko, G.: A Role of Ontology in Information Systems for Support of Scientific and Production Activity. New Trends in Software Methodologies, Tools, and Techniques. Proc. of the eighth SoMeT_09. Fujita, H., Marik, V. (eds), pp. 413-427. IOS Press, Amsterdam (2009)

[20] Боровикова, О.И.: Разработка онтологии для археологического портала знаний. Тр. X Межд. конф. «Проблемы управления и моделирования в сложных системах» (Самара, 23–25 июня 2008 г.) Под ред.: Федосова, Е.А., Кузнецова, Н.А., Виттиха, В.А., сс. 464-470. Самара: Самарский Научный Центр РАН (2008)

[21] Брагинская, Л.П., Григорюк, А.П., Загорулько, Г.Б., Ковалевский, В.В.: Систематизация научных знаний по активной сейсмологии на основе онтологий. Материалы IV Межд. конф. «Современные информационные технологии для научных исследований в области наук о Земле. ITES-

2016» (Южно-Сахалинск, 7–11 августа 2016). Труды конференции, сс. 70-71 (2016)

[22] Гамма, Э., Хелм, Р., Джонсон, Р., Влиссидес, Дж.: Приемы объектно-ориентированного проектирования. Паттерны проектирования. СПб.: «Питер» (2001)

[23] Загорулько, Г.Б.: Разработка онтологии для интернет-ресурса поддержки принятия решений в слабоформализованных областях. Онтология проектирования, 6 (4), сс. 485-500 (2016)

[24] Загорулько, Г.Б., Молородов, Ю.И., Федотов, А.М.: Систематизация знаний по теплофизическим свойствам веществ. Вестник Новосибирского государственного университета. Серия: Информационные технологии, 12 (3), сс. 48-56 (2014)

[25] Загорулько, Ю.А., Боровикова, О.И., Загорулько, Г.Б.: Организация содержательного доступа к гуманитарным информационным ресурсам на основе онтологий. Электронные библиотеки: перспективные методы и технологии, электронные коллекции. Тр. 9-й Всерос. науч. конф. RCDL'2007, 1, сс. 217-224. Переславль-Залесский: Изд-во «Университет города Переславля» (2007)

[26] Загорулько, Ю.А., Загорулько, Г.Б., Боровикова, О.И.: Технология создания тематических интеллектуальных научных интернет-ресурсов, базирующаяся на онтологии. Программная инженерия, (2), сс. 51-60 (2016)

[27] Загорулько, Ю.А., Загорулько, Г.Б.: Использование онтологий в экспертных системах и системах поддержки принятия решений. Труды Второго симпозиума «Онтологическое моделирование» (Казань, 11–12 октября 2010 г.), сс. 321-354. М.: ИПИ РАН (2011)

[28] Нариньяни, А.С.: Недоопределенность в системах представления и обработки знаний. Известия АН СССР. Техническая кибернетика, (5), сс. 3-28 (1986)

[29] Портал Ассоциации ODPA (Association for Ontology Design & Patterns). http:// ontologydesignpatterns.org

[30] NeOn Project. http://www.neon-project.org

[31] Скворцов, Н.А., Калиниченко, Л.А., Ковалев, Д.Ю.: Концептуальное моделирование предметных областей с интенсивным использованием данных. Аналитика и управление данными в областях с интенсивным использованием данных: XVIII Межд. конф. DAMDID / RCDL'2016 (11–14 октября 2016 года, Ершово, Московская обл., Россия): труды конференции. Под ред. Калиниченко, Л.А., Манолопулоса, Я., Кузнецова, С.О., сс. 34-42. М.: ФИЦ ИУ РАН (2016)

[32] Соколова, Е.Г., Кононенко, И.С., Загорулько, Ю.А.: Проблемы описания компьютерной лингвистики в виде онтологии для портала знаний. Компьютерная лингвистика и интеллектуальные технологии: По материалам ежегодной Межд. конф. «Диалог» (Бекасово, 4–8 июня 2008 г.), 7 (14), сс. 482-487. М.: РГГУ (2008)

Ontological description of meteorological and climate data collections

© A.A. Bart

Tomsk State University,
Tomsk, Russia

bart@math.tsu.ru

© A.Z. Fazliev © A.I. Privezentsev © E.P. Gordov © I.G. Okladnikov © A.G.Titov

Institute of Atmospheric Optics SB RAS,
Tomsk, Russia

faz@iao.ru remake@iao.ru gordov@scert.ru oig@scert.ru titov@scert.ru

Abstract. The first version of the primitive OWL-ontology of collections of climate and meteorological data of Institute of Monitoring of Climatic and Ecological Systems SB RAS is presented. The ontology is a component of expert and decision-making support systems intended for quick search for climate and meteorological data suitable for solution of a certain class of applied problems.

Keywords: ontology description of object domains, systematization of domain data, climate and meteorological data.

1 Introduction

Today every large meteorological center uses original meteorological models for calculation of climate and meteorological parameters, which can differ both in the level of detail and set of calculated values of physical parameters. During the reanalysis of a meteorological situation, key meteorological parameters corresponding to measurements at weather stations are usually taken into account.

The results of climatic numerical simulation, weather forecast, or reanalysis of meteorological fields are collections of meteorological parameters that characterize the state of the atmosphere. They are represented by data arrays in common formats, e.g., grib [7], netCDF [12], HDF5 [8], etc.

At Institute of Monitoring of Climate and Ecological Systems SB RAS (IMCES SB RAS), the data processing environment [3] has been developed for representing collections of meteorological data; the environment is provided by sets of metadata that characterize physical parameters entering into the above collections. The practice showed restriction of the use of only localized applications in this environment. Inclusion of external applications resulted in creation of a new system – virtual information platform "Climate+" [17], where data are represented in the netCDF format.

When using climate data from different collections of numerous data manufacturers, the problem arises of ambiguous identification of physical parameters from these collections. The sense of physical parameters in the collections agrees with physical parameters advised by World Meteorological Organization (WMO). They are described in the taxonomy of the WMO ontology Codes Registry [19], as well as in the taxonomy of the ontology of the GRIB Discipline Collection [16] intended for the use in the Climate Information Platform for Copernicus (CLIPC).

The ontology description of data collections in the form of a primitive (simplified) formal OWL-ontology is intended for the selection of data collections within an expert system, which can be used during solution of an applied task of an object domain.

The ontology approach selected for the solution of the problem stated consists in the following. An ontology description is constructed for an applied problem. In addition to the physical statement, the description should include the mathematical statement of the task, i.e., a mathematical model with equations. Variables, which conform the WMO classification, and limitations are described in the form of an OWL-ontology. On the one hand, the set of parameters includes common meteorological parameters, such as sea level pressure, surface pressure, air temperature and humidity, wind speed and direction, and so on. This allows comparison of the computed values with the weather station measurement results. On the other hand, both meteorological and climate models supplemented by an applied task compose a component of a more complex model, where the results of prognostic calculations by climate/meteorological parameters are used for the solution of applied problems in different fields of human activity. This, in turn, enriches collections of climate and meteorological data with values of new physical parameters.

2 Virtual data processing environment

Approaches used in the creation of the prototype of a

Proceedings of the XIX International Conference "Data Analytics and Management in Data Intensive Domains" (DAMDID/RCDL'2017), Moscow, Russia, October 10-13, 2017

subject virtual data processing environment (VDPS) for the analysis, estimation, and forecast of the impacts of global climate changes on the natural environment and climate of a region were mainly developed during the design of the "Climate" web GDS [4,5]. This subject GDS has been designed with the use of up-to-date information and communication technologies, is based on the conceptions of spatial data infrastructure (SDI) [2, 10], and grounds a software infrastructure for the complex use of geophysical data and information support of integrated multidisciplinary scientific researches in the modern quantitative meteorology. We have selected it as a subject component of VDPS for Earth sciences. A web geoportal [1, 9] is a single access point to subject spatial data, processing procedures and results [1, 9]. The portal allows a user to search for geoinformation resources in metadata catalogues, to form samples of spatial data according to their characteristics (access functionality), and to manage tools and applications for data processing and mapping.

The GDS Web Client [6, 13] is the main tool of the user's desktop. It ensures the fulfillment of OGC requirements for web services: spatial data visualization (Web Map Service—WMS), data representation in vector (Web Feature Service—WFS) and bitmap formats (Web Coverage Service—WCS), and their geospatial processing. It provides for the access to collections of climate data and tools for their analysis and visualization of the results via typical GDS graphical web browser. The Web Client satisfies the general requirements of INSPIRE standards and allows selection of data set, processing type, geographic region for the analysis of processes, and representation of the processing results of spatial data sets in the form of WMS/WFS map layers in bitmap (PNG, JPG, GeoTIFF), vector (KML, GML, Shape), and binary formats (NetCDF).

Today, the VDPS prototype combines data collections (reanalyses and climate simulation results and weather station measurements) within the unified geoportal, supports the statistical analysis of archive and required data, and provides access to the WRF and «Planet Simulator» models. In particular, a user can run a VDPS-integrated model, preprocess the results, process them numerically and analyze, and gain the results in graphical representation. The prototype provides for specialists that participate in a multidisciplinary research process prompt tools for integral study of climate and ecological systems on the global and regional scales. With these tools, a user that does not know programming is able of processing and graphically representing multidimensional observation and simulation data in the unified interface with the use of the web browser.

3 VDPS prototype capabilities

Support of the following data sets is built in the prototype: NCEP/NCAR reanalysis, ed. II, JMA/CRIEPI JRA-25 reanalysis, ECMWF ERA-40 reanalysis, ECMWF ERA Interim, MRI/JMA APHRODITE's Water Resources Project data, DWD Global Precipitation Climatology Centre data, GMAO Modern Era-Retrospective analysis for Research and Applications (MERRA), reanalysis of the joint Project «Monitoring atmospheric composition and climate (MACC)», NOAA-CIRES Twentieth Century Global Reanalysis, ver. II, NCEP Climate Forecast System Reanalysis (CFSR), simulation results obtained with the use of global and regional climate and meteorological models. Observation data from weather stations from the territory of the former USSR for the 20th century included in the PostGIS database are also accessible.

Data processing

1. Statistical characteristics of meteorological parameters: sample mean, variance, excess, median, minimum and maximum, and asymmetry.
2. Derived climate parameters: vegetation period duration, sum of effective temperature, Selyaninov hydrothermal coefficient.
3. Periodic variations: standard deviation, norms, aberrations, amplitudes of diurnal and annual variations.
4. Non-periodic variations: duration and repeatability of atmospheric phenomena with meteorological parameters below or above the limits specified at different time points.

Then a user can either analyze the results or continue adding new layers on the map. To study the results, the user is provided for a possibility of selecting a geographical region, scaling, getting values from all layers at a point, additionally processing earlier results (e.g., comparison between data from different layers). In addition to the direct analysis of geophysical data, a used can carry out joint researches with other user, share the results, and use proper data collections in the processing. In general, this hardware-software complex provides for distributed access, processing and visualization of large collections of geospatial data with the use of cloud technologies.

The data processing environment "Climate" developed at IMCES SB RAS limits possibilities of users by local software applications. A current task is to extend the environment by external user applications. For this, the corresponding problems should be specified in general. Below we describe one of possible classes of problems connected with decision-making.

4 General definition of the problem

The "Climate+" virtual information platform includes collections of meteorological and climate data. It is intended for the data representation with the use of GIS technologies. Its further development is oriented to providing researchers possibilities of using selected data sets or their parts as input data. Most collections include data related to some (not all) spatiotemporal objects of the Earth; different collections often include different sets of physical parameters. To search for required spatiotemporal objects and their meteorological and clime characteristics, it was necessary to create a corresponding expert system on the basis of a knowledge base on spatial objects of the data collections and their parameters.

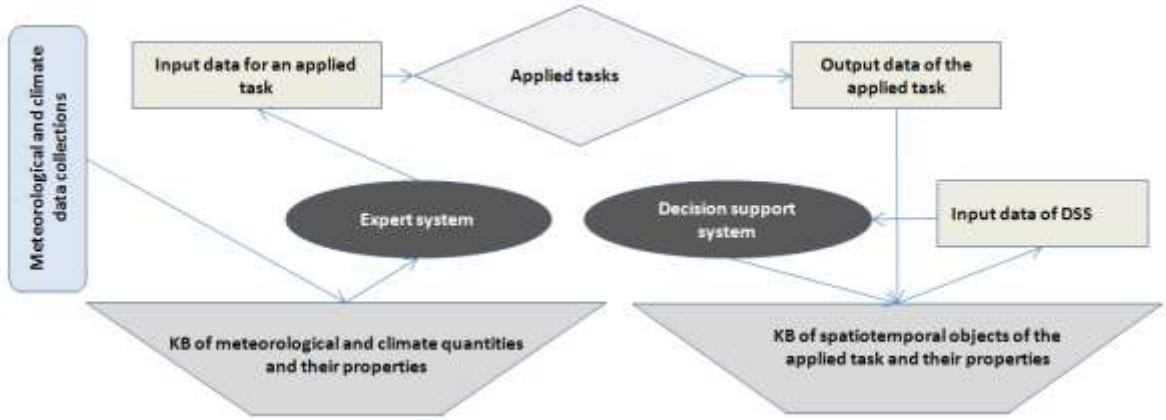


Figure 1 Simplified block-diagram of "Climate +" platform modification

Figure 1 shows a simplified block-diagram which is a basis of the "Climate +" platform modification. There are three groups of subsystems: meteorological and climate data collections; subsystem for work with knowledge bases (expert system for selecting input data for applied tasks and decision-making support system), and applied tasks with their input and output data. The data representation services are omitted.

In this work, we discuss questions of creation of a knowledge base for the expert system. The main problem which has been solved is substantiation of the reduction problem solution [20] or, in other words, construction of typical individuals of an OWL-ontology that characterize properties of spatiotemporal objects from the collections. The development of the conceptual part of the ontology (T- and R-box) is connected in our solution with classification of meteorological and climate parameters and is briefly described below.

5 Taxonomy of meteorological parameters

The OWL DL language [14] is used for the ontology description of the domain that generalizes, in particular, related spatiotemporal objects. These objects can be an air layer over a bounded territory, upper soil layer on this territory, or, in more specific cases, forests, fields, or long roads. There are physical and chemical processes connected with the objects; they are described by numerical models and used in calculations. Input values of the physical parameters are required for the calculations. The processes under study can relate to different temporal and spatial scales and be described on different levels of detail. Let us note that coupling of several mathematical models requires knowledge of sets of input and output parameters and their spatiotemporal characteristics.

The taxonomy of physical parameters allows forming sets of properties of spatiotemporal objects of a domain for solution of specific applied tasks. This taxonomy is used in the OWL-ontology for T-box construction.

When developing the decision-making support system on the basis of both meteorological and climate data, the parameters should be matched. Therefore, the WMO classification in version [11] is included in the ontology. This matching allows describing applied tasks of the domain in common terms.

There are climatic and meteorological resources [16, 19] that use the WMO classification of names of meteoparameters for the GRIB format for data storage [7]. First of all, WMO Codes Registry created for the aviation with the aim of supporting data exchange in the AvXML format; it is based on RDF and SKOS recommendation.

In our OWL-ontology of climate information resources, we created classes and individuals that correspond to names of meteorological parameters, e.g., the Meteorological_Products class and subclasses, according to [11]. In the primitive OWL-ontology of climate information resources described below, classes and individuals are created that correspond to names of meteorological parameters according to [11]. Individuals that unambiguously characterize physical parameters by their name [11] have been created in each subclass Thermodynamic_Stability_category, Atmospheric_Chemical_Constituents_category, Electrodynamics_category, Mass_category, Long-wave_radiation_category, Temperature_category, Short-wave_radiation_category, Aerosols_category, Moisture_category, Radiology_Imagery_category, Momentum_category, Trace_Gases_category, Cloud_category, and Physical_Atmospheric_category.

For the INMCM4 collection, which corresponds to output data of the INMCM4 climate model of general atmospheric and ocean circulation [18], classes and subclasses were created corresponding to model variables. These classes agree to the corresponding WMO classes.

6 Primitive ontology of "Climate+" platform data

The OWL DL developed and formalized ontology of climate information resources describes the current state of collections of data arrays of the data processing environment as one of the main Russian information resources on climate data. Numerical data are represented by data arrays that are stored in netCDF files. The data arrays are grouped in data sets. All data arrays

in a set should: (a) be received at one temporal or spatial grid; (b) cover the same time interval; (c) be received under the same simulation or observation conditions (if possible); (d) be represented by a set of netCDF files, which include the same physical parameters. The data sets are grouped in data collections. A data collection is an ensemble of data sets received by an organization within a project, but represented on different spatial or temporal grids or for different model scenarios. In particular, a collection can consists of the only data set.

The basic classes in the OWL-ontology are: Collection, Spatiotemporal_object, Organization, Data_set, Data_array, Scenario, Spatial_resolution, Physical_quantity, Physical_quantity_values, Unit, Longitudes_array, Time_step, Latitudes_array, Height_levels_array, and Times_array. The spatiotemporal system is a four-dimensional object determined by arrays of numerical values of longitudes (Longitudes_array), latitudes (Latitudes_array), height levels (Height_levels_array), and time labels (Times_array), which are subclasses of the class of the list of values of a physical parameter and, therefore, numerical arrays of one physical parameter (Physical_quantity) in certain measurement units (Unit). They can be described by: the number of members of the array of a physical parameter (has_number_of_values), its minimal value (has_minimum_value) and maximal value (has_maximum_value), or by numerical values of the parameter (has_value). A data array (Data_array) is an ordered list of numerical values of a physical parameter (Physical_quantity), as a property of the spatiotemporal system (has_spatiotemporal_system), at each 4D point (longitude, latitude, height level, and time) of the spatiotemporal system (Spatiotemporal_system). In the OWL-ontology, a data array (Data_array) is a subclass of the class Physical_quantity_values and, hence, is a numerical array of values of one physical parameter (Physical_quantity) in certain measurement units (Unit); it is described by the number of members (has_number_of_values), maximal values (has_minimum_value) and minimal values (has_maximum_value) of the physical parameter. A data array (Data_array) belongs (has_data_array) to a data set (Data_set), which differs from other data sets by the model scenario (Scenario), spatial resolution (Spatial_resolution), time step (Time_step), and belonging to one collection (Collection). A data collection (Collection) consists of (has_data_set) data sets (Data_set) and belongs (has_organization) to one organization (Organization). The OWL properties of the climate data ontology are represented in Tables 1 and 2.

Table 1 Object properties of the ontology of climate information resources

Domain	Object Property	Range	id
Collection	has_organization	Organization	o01
Collection	has_data_set	Data_set	o02
Data_set	has_scenario	Scenario	o03
Data_set	has_spatial_resolution	Spatial_resolution	o04
Data_set	has_time_step	Time_step	o05
Data_set	has_data_array	Data_array	o06
Physical_quantity_values	has_physical_quantity	Physical_quantity	o07
Physical_quantity_values	has_unit	Unit	o08
Data_array	has_spatiotemporal_object	Spatiotemporal_object	o09
Spatiotemporal_object	has_longitudes_array	Longitudes_array	o10
Spatiotemporal_object	has_latitudes_array	Latitudes_array	o11
Spatiotemporal_object	has_height_levels_array	Height_levels_array	o12
Spatiotemporal_object	hat_times_array	Times_array	o13

Definitions of object properties are given in first three rows of Table 1; their unique identifying properties, in the fourth row; the range of definition (the first row) and range of values (the third row) are specified for each property. Definition of the data array properties are given in the first three rows of Table 2; unique identifying properties are given in the fourth row. The range of definition (the first row) and range of values (the third row) are specified for each property from the second row.

Table 2 Data type properties in the ontology of climate information resources

Domain	Datatype Property	Range	id
Physical_quantity_values	has_number_of_values	int	d01
Physical_quantity_values	has_minimum_value	float	d02
Physical_quantity_values	has_maximum_value	float	d03
Physical_quantity_values	has_value	float	d04
Times_array	has_time_start	str	d05
Times_array	has_time_end	str	d06

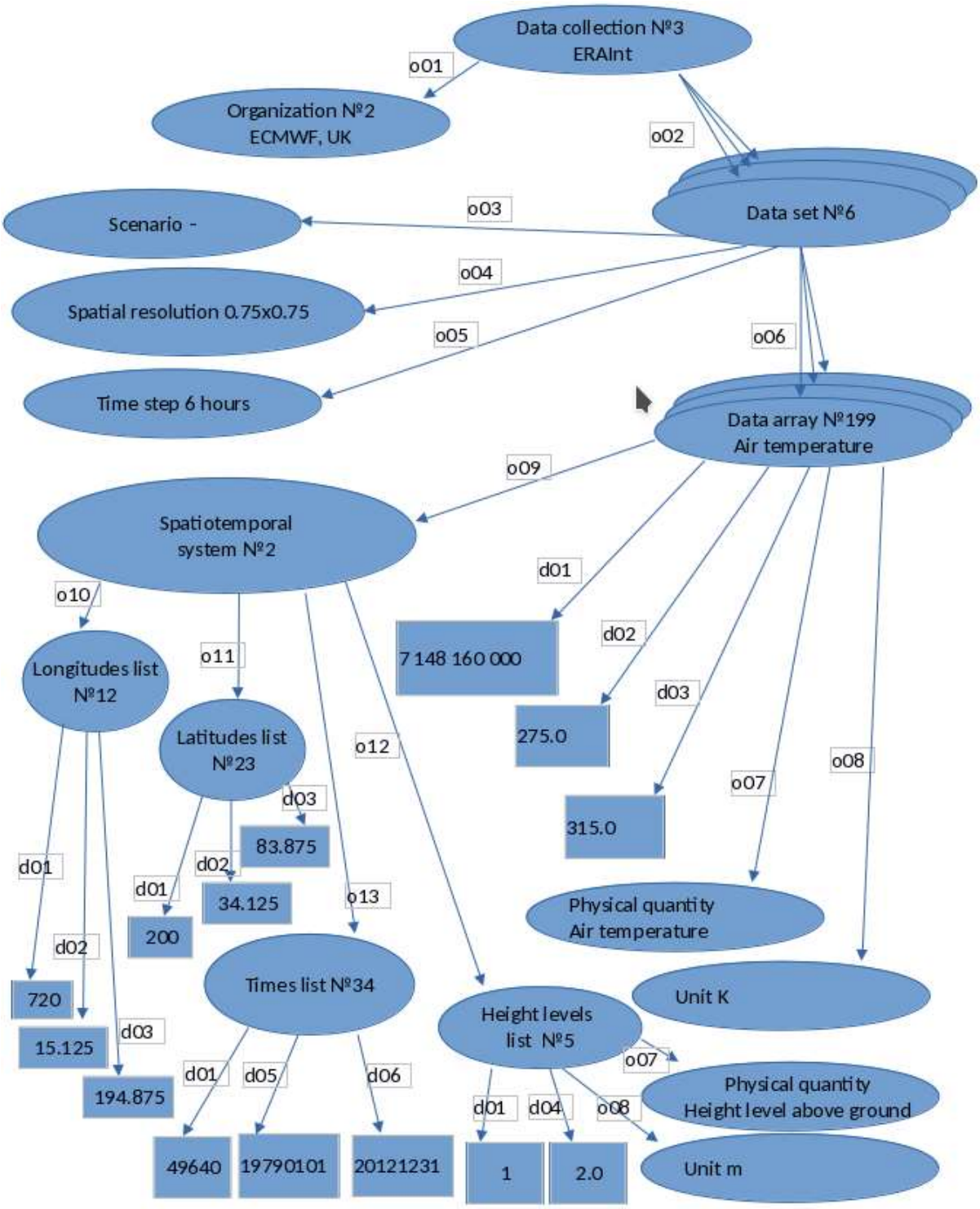


Figure 2 Simplified representation of individual describing ERAInt data collections

Figure 2 exemplifies a simplified individual of the OWL-ontology of climate information resources, used in the description of a ERAInt data collection, within the formal description of RDF resources [15].

Individuals of the OWL-ontology are shown in ovals; literal values are given in rectangles; the arrows show properties with unique identifiers in small rectangles, taken from Tables 1 and 2. Three arrows mean probable property cardinality higher than unity. Three overlapped ovals mean probable number of individuals of the OWL-ontology larger than unity. The individual "Data_collection" is connected by the property "has_data_set" with the individuals "Data_set", each of which is connected by the property "has_data_array" with individuals "Data_array".

The domain analysis of climate numerical data arrays of the "Climate+" platform, stored as NetCDF files, allows the description of a primitive ontology of climate data of this platform in the OWL DL language. The primitive ontology is a simple and easily extended systematization of information resources required for the further work on the development of the decision-making support system.

To construct the climate data ontology of the "Climate+" platform the software has been developed for the formation of the fact-based block (A-box). An A-box has been formed for the climate data ontology using this software. Facts have been retrieved from the analysis of 80 Tb of climate data from the "Climate+" platform over 13 numerical data collections, which include 36 data sets and 793 data arrays. All the climate data collections include description of 170 spatiotemporal systems and 156 physical parameters that characterize properties of these systems.

7 Conclusions

The prototype of subject virtual data processing environment has been developed to provide for researchers, specialists, and people that make decisions an access to different geographically distributed and georeferenced resources and climate data processing services via a typical web browser. It includes a geoportal, systems for distributed storage, processing, and providing of spatial data and results of their processing. In particular, it allows the simultaneous analysis of several subject sets of climate data with the use of up-to-date statistical methods and, thus, revealing the impacts of climate changes on ecological processes and human activity. After finishing the work on the prototype, different interactive web tools are to be developed for the profound analysis of climatic variables and their derivatives provided by the subject geoportal.

The developed software is used for processing spatial datasets, including observation and reanalysis data, for the spatiotemporal analysis of recent and probable climate changes, with the special focus on extreme climate phenomena in northern latitudes.

The primitive OWL-ontology of climate and meteorological collections of IMCES SB RAS is constructed; it can be used for the search and selection of data for classes of applied problems in coupled decision support systems. The matching of physical parameters of applied tasks with IMCES SB RAS collections is carried out in WMO accepted terms.

8 Acknowledgment

The authors thank the Russian Science Foundation for the support of this work under the grant No16-19-10257.

References

[1] Becirspahic, L. and Karabegovic A.: Web portals for visualizing and searching spatial data. Inform. Comm. Techn., Electr. and Microelectr. (MIPRO), 2015, 38-th International Convention on, Opatija, 2015, 305-311. doi: 10.1109/MIPRO.2015.7160284

[2] Frans J.M., van der Wel.: Spatial data infrastructure for meteorological and climatic data. Meteorol. Appls. 2005. V. 12. No. 1. P. 7-8

[3] Gordov E. P., Okladnikov I. G., Titov A. G.: Application of Web Mapping Technologies for Development of Information-Computational Systems for Georeferenced Data Analysis, Vestnik NGU, Ser. Information Technologies, 9(4), 94-102 (2011). (in Russian)

[4] Gordov E.P., Lykosov V.N., Krupchatnikov V.N., Okladnikov I.G., Titov A.G., Shulgina T.M.: Computational-information technologies for monitoring and modeling of climate change and its consequences. Novosibirsk: Nauka, 199 (2013) (in Russian)

[5] Gordov E.P., Okladnikov I.G., Titov A.G. : Information and computing Web-system for interactive analysis of georeferenced climatic data sets, Vestnik NGU, Ser. Information Technologies, 14(1), 13–22 2016. (in Russian)

[6] Gordov, E., Shiklomanov, A., Okladnikov, I., Prusevich, A., and Titov, A.: Development of Distributed Research Center for analysis of regional climatic and environmental changes, IOP Conf. Series: Earth and Environmental Science, V.48. 012033 (2016)

[7] Guide to the WMO Table Driven Code Form Used for the Representation and Exchange of Regularly Spaced Data In Binary Form: FM 92 GRIB Edition 2 // World Meteorological Organization Extranet. 2003. URL: http://www.wmo.int/pages/prog/www/WMOCodes/Guides/GRIB/GRIB2_062006.pdf

[8] HDF Group - HDF5: https://support.hdfgroup.org/HDF5/

[9] Koshkarev A.V. : Geoportal as a tool to control geospatial data and services, Geospatial data, 2, 6–14 (2008). (in Russian)

[10] Koshkarev A.V., Ryakhovskii A.V., Serebryakov V.A.: Infrastructure of Distributed Environment of Spatial Data Storage, Search and

Processing, Open Education, 5, 61-73 (2010). (in Russian)

[11] NCEP/NCO Production Management Branch. NCEP WMO GRIB2 Documentation // National Weather Service Organization NCEP Central Operations. 2005. http://www.nco.ncep.noaa.gov/ pmb/docs/grib2/grib2_doc.shtml

[12] Network Common Data Form (NetCDF). https://www.unidata.ucar.edu/software/netcdf/

[13] Okladnikov I.G., Gordov E.P., Titov A.G.: Development of climate data storage and processing model. IOP Conf. Series: Earth and Environmental Science, 48, 012030 (2016)

[14] OWL 2 Web Ontology Language. RDF-Based Semantics (Second Edition), Eds: M. Schneider, F.J. Carroll, I. Herman, P.F. Patel-Schneider. W3C Recommendation 11 December 2012, http://www.w3.org/TR/2012/ REC-owl2-rdf-based-semantics-20121211/

[15] Resource Description Framework (RDF): Concepts and Abstract Syntax, W3C Recommendation 10 February 2004, Eds: Graham Klyne, Jeremy J. Carroll, http://www.w3.org/TR/2004/REC-rdf-concepts-20040210/

[16] The GRIB Discipline Collection: [site]. [2004]. URL: http://vocab-test.ceda.ac.uk/collection/ grib/Discipline

[17] Titov A.G., Gordov E.P., Okladnikov I.G.: Hardware-Software Platform «CLIMATE» as a Basis for Local Spacial Data Infrastructure Geoportal, Vestnik NGU, Ser. Information Technologies, 10(4), 104-111 (2012).) (in Russian)

[18] Volodin E. M., Dianskii N. A., Gusev A. V.: Simulating present-day climate with the INMCM 4.0 coupled model of the atmospheric and oceanic general circulations, Izvestiya, Atmospheric and Oceanic Physics, 46(4), 414-431 (2010)

[19] WMO Codes Registry: (2013). URL: http://codes.wmo.int/grib2

[20] Zinov'ev A. A.: Foundations of the logical theory of scientific knowledge (Complex Logic), D. Reidel Publishing Company. 264 P

Организация экспериментов в ОИИД

Organization of experiments in data intensive research

Data Mining Design and Systematic Modelling

© Yannic Kropp © Bernhard Thalheim

Christian Albrechts University Kiel, Department of Computer Science,
D-24098 Kiel, Germany

yk@is.informatik.uni-kiel.de thalheim@is.informatik.uni-kiel.de

Abstract. Data mining is currently a well-established technique and supported by many algorithms. It is dependent on the data on hand, on properties of the algorithms, on the technology developed so far, and on the expectations and limits to be applied. It must be thus matured, predictable, optimisable, evolving, adaptable and well-founded similar to mathematics and SPICE/CMM-based software engineering. Data mining must therefore be systematic if the results have to be fit to its purpose. One basis of this systematic approach is model management and model reasoning. We claim that systematic data mining is nothing else than systematic modelling. The main notion is the notion of the model in a variety of forms, abstraction and associations among models.

Keywords: data mining, modelling, models, framework, deep model, normal model, modelling matrix.

1 Introduction

Data mining and analysis is nowadays well-understood from the algorithms side. There are thousands of algorithms that have been proposed. The number of success stories is overwhelming and has caused the big data hype. At the same time, brute-force application of algorithms is still the standard. Nowadays data analysis and data mining algorithms are still taken for granted. They transform data sets and hypotheses into conclusions. For instance, cluster algorithms check on given data sets and for a clustering requirements portfolio whether this portfolio can be supported and provide as a set of clusters in the positive case as an output. The Hopkins index is one of the criteria that allow to judge whether clusters exist within a data set. A systematic approach to data mining has already been proposed in [3, 17]. It is based on mathematics and mathematical statistics and thus able to handle errors, biases and configuration of data mining as well. Our experience in large data mining projects in archaeology, ecology, climate research, medical research etc. has however shown that ad-hoc and brute-force mining is still the main approach. The results are taken for granted and believed despite the modelling, understanding, flow of work and data handling pitfalls. So, the results often become dubious.

Data are the main source for information in data mining and analysis. Their quality properties have been neglected for a long time. At the same time, modern data management allows to handle these problems. In [16] we compare the critical findings or pitfalls of [21] with resolution techniques that can be applied to overcome the crucial pitfalls of data mining in environmental sciences reported there. The algorithms themselves are another source of pitfalls that are typically used for the solution of data mining and analysis tasks. It is neglected that an algorithm also has an application area, application restrictions, data requirements, results at certain granularity and precision. These problems must be systematically tackled if we want to rely on the results of mining and analysis. Otherwise analysis may become misleading, biased, or not possible. Therefore, we explicitly treat properties of mining and analysis. A similar observation can be made for data handling.

Data mining is often considered to be a separate sub-discipline of computer engineering and science. The statistics basis of data mining is well accepted. We typically start with a general (or better generic) model and use for refinement or improvement of the model the data that are on hand and that seem to be appropriate. This technique is known in sciences under several names such as inverse modelling, generic modelling, pattern-based reasoning, (inductive) learning, universal application, and systematic modelling.

Data mining is typically not only based on one model but rather on a model ensemble or model suite The association among models in a model suite is explicitly specified. These associations provide an explicit form via model suites. Reasoning techniques combine methods from logics (deductive, inductive, abductive, counter-inductive, etc.), from artificial intelligence (hypothetic, qualitative, concept-based, adductive, etc.), computational methods (algorithmics [6], topology, geometry, reduction, etc.), and cognition (problem representation and solving, causal reasoning, etc.).

These choices and handling approaches need a systematic underpinning. Techniques from artificial intelligence, statistics, and engineering are bundled within the CRISP framework (e.g. [3]). They can be enhanced by techniques that have originally been developed for modelling, for design science, business informatics, learning theory, action theory etc.

We combine and generalize the CRISP, heuristics, modelling theory, design science, business informatics, statistics, and learning approaches in this paper. First, we introduce our notion of the model. Next we show how

Proceedings of the XIX International Conference "Data Analytics and Management in Data Intensive Domains" (DAMDID/RCDL'2017), Moscow, Russia, October 10–13, 2017

data mining can be designed. We apply this investigation to systematic modelling and later to systematic data mining. It is our goal to develop a holistic and systematic framework for data mining and analysis. Many issues are left out of the scope of this paper such as a literature review, a formal introduction of the approach, and a detailed discussion of data mining application cases.

2 Models and Modelling

Models are principle instruments in mathematics, data analysis, modern computer engineering (CE), teaching any kind of computer technology, and also modern computer science (CS). They are built, applied, revised and manufactured in many CE&CS sub-disciplines in a large variety of application cases with different purposes and context for different communities of practice. It is now well understood that models are something different from theories. They are often intuitive, visualizable, and ideally capture the essence of an understanding within some community of practice and some context. At the same time, they are limited in scope, context and the applicability.

2.1 The Notion of the Model

There is however a general notion of a model and of a conception of the model: *A* **model** *is a well-formed, adequate, and dependable instrument that represents origins* [9, 29, 30].

Its criteria of well-formedness, adequacy, and dependability must be commonly accepted by its community of practice within some context and correspond to the functions that a model fulfills in utilization scenarios.

A well-formed instrument is *adequate* for a collection of origins if it is *analogous* to the origins to be represented according to some analogy criterion, it is more *focused* (e.g. simpler, truncated, more abstract or reduced) than the origins being modelled, and it sufficiently satisfies its *purpose*.

Well-formedness enables an instrument to be *justified* by an *empirical corroboration* according to its objectives, by *rational coherence* and *conformity* explicitly stated through conformity formulas or statements, by *falsifiability* or *validation*, and by *stability* and *plasticity* within a collection of origins.

The instrument is *sufficient* by its *quality* characterization for internal quality, external quality and quality in use or through quality characteristics [28] such as correctness, generality, usefulness, comprehensibility, parsimony, robustness, novelty etc. Sufficiency is typically combined with some *assurance evaluation* (tolerance, modality, confidence, and restrictions).

2.2 Generic and Specific Models

The general notion of a model covers all aspects of adequateness, dependability, well-formedness, scenario, functions and purposes, backgrounds (grounding and basis), and outer directives (context and community of practice). It covers all known so far notions in agriculture, archaeology, arts, biology, chemistry, computer science, economics, electro-technics, environmental sciences, farming, geosciences, historical sciences, languages, mathematics, medicine, ocean sciences, pedagogical science, philosophy, physics, political sciences, sociology, and sports. The models used in these disciplines are instruments used in certain scenarios.

Sciences distinguish between general, particular and specific things. Particular things are specific for general things and general for specific things. The same abstraction may be used for modelling. We may start with a general model. So far, nobody knows how to define general models for most utilization scenarios. Models *function* as *instruments* or tools. Typically, instruments come in a variety of forms and fulfill many different functions. Instruments are partially independent or autonomous of the thing they operate on. Models are however special instruments. They are used with a specific intention within a utilization scenario. The quality of a model becomes apparent in the context of this scenario.

It might thus be better to start with generic models. A ***generic model*** [4, 26, 31, 32] is a model which broadly satisfies the purpose and broadly functions in the given utilization scenario. It is later tailored to suit the particular purpose and function. It generally represents origins of interest, provides means to establish adequacy and dependability of the model, and establishes focus and scope of the model. Generic models should satisfy at least five properties:(i) they must be accurate; (ii) the quality of generic models allows that they are used consciously; (iii) they should be descriptive, not evaluative;(iv) they should be flexible so that they can be modified from time to time;(v) they can be used as a first "best guess".

2.3 Model Suites

Most disciplines integrate a variety of models or a *society of models*, e.g. [7, 14] Models used in CE&CS are mainly at the same level of abstraction. It is already well-known for threescore years that they form a *model ensemble* (e.g. [10, 23]) or horizontal *model suite* (e.g. [8, 27]). Developed models vary in their scopes, aspects, and facets they represent and their abstraction.

A ***model suite*** consists of a set of models $\{M_1,...,M_n\}$, of an association or collaboration schema among the models, of controllers that maintain consistency or coherence of the model suite, of application schemata for explicit maintenance and evolution of the model suite, and of tracers for the establishment of the coherence.

Multi-modelling [11, 19, 24] became a culture in CE&CS. Maintenance of coherence, co-evolution, and consistency among models has become a bottleneck in development. Moreover, different languages with different capabilities have become an obstacle similar to multi-language retrieval [20] and impedance mismatches. Models are often loosely coupled. Their dependence and relationship is often not explicitly expressed. This problem becomes more complex if models are used for different purposes such as

construction of systems, verification, optimization, explanation, and documentation.

2.4 Stepwise Refinement of Models

Refinement of a model to a particular or special model provides mechanisms for model transformation along the adequacy, the justification and the sufficiency of a model. Refinement is based on *specialization* for better suitability of a model, on *removal* of unessential elements, on *combination* of models to provide a more holistic view, on *integration* that is based on binding of model components to other components and on *enhancement* that typically improves a model to become more adequate or dependable.

Control of correctness of refinement [33] for information systems takes into account (A) a focus on the refined structure and refined vocabulary, (B) a focus to information systems structures of interest, (C) abstract information systems computation segments, (D) a description of database segments of interest, and

(E) an equivalence relation among those data of interest.

2.5 Deep Models and the Modelling Matrix

Model development is typically based on an explicit and rather quick description of the 'surface' or *normal model* and on the mostly unconditional acceptance of a *deep model*. The latter one directs the modelling process and the surface or normal model. Modelling itself is often understood as development and design of the normal model. The deep model is taken for granted and accepted for a number of normal models.

The deep model can be understood as the common basis for a number of models. It consists of the grounding for modelling (paradigms, postulates, restrictions, theories, culture, foundations, conventions, authorities), the outer directives (context and community of practice), and basis (assumptions, general concept space, practices, language as carrier, thought community and thought style, methodology, pattern, routines, commonsense) of modelling. It uses a collection of undisputable elements of the background as grounding and additionally a disputable and adjustable basis which is commonly accepted in the given context by the community of practice. Education on modelling starts, for instance, directly with the deep model. In this case, the deep model has to be accepted and is thus hidden and latent.

A (modelling) matrix is something within or from which something else originates, develops, or takes from. The matrix is assumed to be correct for normal models. It consists of the deep model and the modelling scenarios. The modelling *agenda* is derived from the modelling scenario and the utilization scenarios. The modelling scenario and the deep model serve as a part of the *definitional frame* within a model development process. They define also the capacity and potential of a model whenever it is utilized.

Deep models and the modelling matrix also define some frame for adequacy and dependability. This frame is enhanced for specific normal models. It is then used for a statement in which cases a normal model represents the origins under consideration.

2.6 Deep Models and Matrices in Archaeology

Let us consider an application case. The CRC *1266*[1] *"Scales of Transformation – Human Environmental Interaction in Prehistoric and Archaic Societies"* investigates processes of transformation from 15,000 BCE to 1 BCE, including crisis and collapse, on different scales and dimensions, and as involving different types of groups, societies, and social formations. It is based on the matrix and a deep model as sketched in Figure 1. This matrix determines which normal models can still be considered and which not. The initial model for any normal model accepts this matrix.

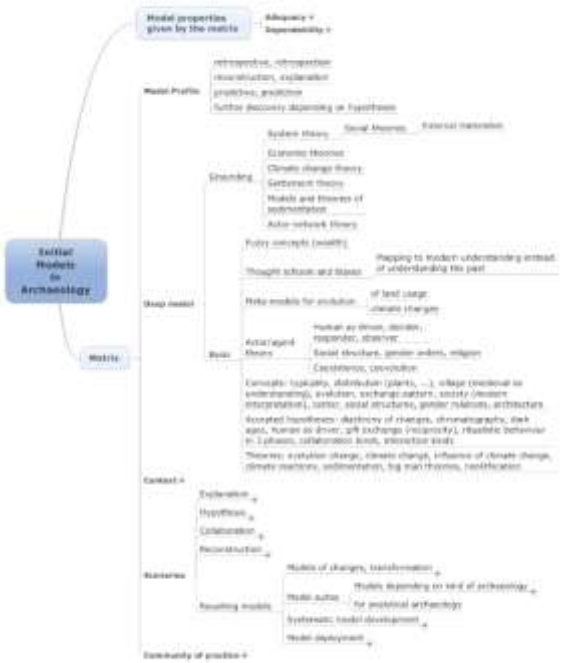


Figure 1 Modeling in archaeology with a matrix

We base our consideration on the matrix and the deep model on [19] and the discussions in the CRC. Whether the deep model or the model matrix is appropriate has already been discussed. The final version presented in this paper illustrates our understanding.

2.7 Stereotyping of a Data Mining Process

Typical modeling (and data mining) processes follow some kind of ritual or typical guideline, i. e. they are stereotyped. The *stereotype* of a modelling process is based on a general modelling situation. Most modelling methodologies are bound to one stereotype and one kind

[1]https://www.sfb1266.uni-kiel.de/en

of model within one model utilization scenario. Stereotypes are governing, conditioning, steering and guiding the model development. They determine the model kind, the background and way of modelling activities. They persuade the activities of modelling. They provide a means for considering the economics of modelling. Often, stereotypes use a definitional frame that primes and orients the processes and that considers the community of practice or actors within the model development and utilization processes, the deep model or the matrix with its specific language and model basis, and the agenda for model development. It might be enhanced by initial models which are derived from generic models in accordance to the matrix.

The model utilization scenario determines the function that a model might have and therefore also the goals and purposes of a model.

2.8 The *Agenda*

The agenda is something like a guideline for modeling activities and for model associations within a model suite. It improves the quality of model outcomes by spending some effort to decide what and how much reasoning to do as opposed to what activities to do. It balances resources between the data-level actions and the reasoning actions. E.g. [17] uses an agent approach with preparation agents, exploration agents, descriptive agents, and predictive agents. The agenda for a model suite uses thus decisions points that require agenda control according to performance and resource considerations. This understanding supports introspective monitoring about performance for the data mining process, coordinated control of the entire mining process, and coordinated refinement of the models. Such kind of control is already necessary due to the problem space, the limitations of resources, and the amount of uncertainty in knowledge, concepts, data, and the environment.

3 Data Mining Design

3.1 Conceptualization of Data Mining and Analysis

The data mining and analysis task must be enhanced by an explicit treatment of the languages used for concepts and hypotheses, and by an explicit description of knowledge that can be used. The algorithmic solution of the task is based on knowledge on algorithms that are used and on data that are available and that are required for the application of the algorithms. Typically, analysis algorithms are iterative and can run forever. We are interested only in convergent ones and thus need termination criteria. Therefore, conceptualization of the data mining and analysis task consists of a detailed description of *six main parameters* (e.g. for inductive learning [34]):

(a) The *data analysis algorithm*: Algorithm development is the main activity in data mining research. Each of these algorithms transfers data and some specific parameters of the algorithm to a result.

(b) The *concept space*: the concept space defines the concepts under consideration for analysis based on certain language and common understanding.

(c) The *data space*: The data space typically consists of a multi-layered data set of different granularity. Data sets may be enhanced by metadata that characterize the data sets and associate the data sets to other data sets.

(d) The *hypotheses space*: An algorithm is supposed to map evidence on the concepts to be supported or rejected into hypotheses about it.

(e) The *prior knowledge space*: Specifying the hypothesis space already provides some prior knowledge. In particular, the analysis task starts with the assumption that the target concept is representable in a certain way.

(f) The *acceptability and success criteria*: Criteria for successful analysis allow to derive termination criteria for the data analysis.

Each instantiation and refinement of the six parameters leads to specific data mining tasks.

The result of data mining and data analysis is described within the knowledge space. The data mining and analysis task may thus be considered to be a transformation of data sets, concept sets and hypothesis sets into chunks of knowledge through the application of algorithms.

Problem solving and modelling considers, however, typically six aspects [16]:

(1) *Application, problems, and users*: The domain consists of a model of the application, a specification of problems under consideration, of tasks that are issued, and of profiles of users.

(2) *Context*: The context of a problem is anything what could support the problem solution, e.g. the sciences' background, theories, knowledge, foundations, and concepts to be used for problem specification, problem background, and solutions.

(3) *Technology*: Technology is the enabler and defines the methodology. It provides [23] means for the flow of problem solving steps, the flow of activities, the distribution, the collaboration, and the exchange.

(4) *Techniques and methods*: Techniques and methods can be given as algorithms. Specific algorithms are data improvers and cleaners, data aggregators, data integrators, controllers, checkers, acceptance determiners, and termination algorithms.

(5) *Data*: Data have their own structuring, their quality and their life span. They are typically enhanced by metadata. Data management is a central element of most problem solving processes.

(6) *Solutions*: The solutions to problem solving can be formally given, illustrated by visual means, and presented by models. Models are typically only normal models. The deep model and the matrix is already provided by the context and accepted by the community of practice in dependence of the needs of this community for the given application scenario. Therefore, models

may be the final result of a data mining and analysis process beside other means.

Comparing these six spaces with the six parameters we discover that only four spaces are considered so far in data mining. We miss the user and application space as well as the representation space. Figure 2 shows the difference.

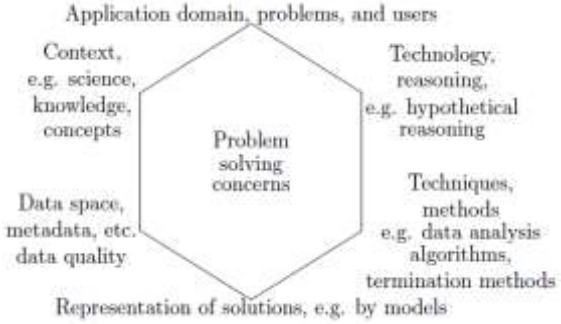


Figure 2 Parameters of Data Mining and the Problem Solving Aspects

3.2 Meta-models of Data Mining

An abstraction layer approach separates the application domain, the model domain and the data domain [17]. This separation is illustrated in Figure 3.

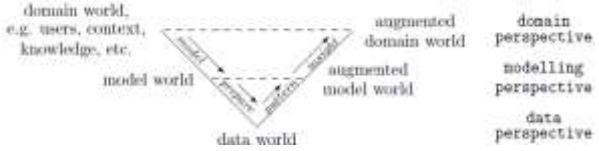


Figure 3 The V meta-model of Data Mining Design

The data mining design framework uses the inverse modeling approach. It starts with the consideration of the application domain and develops models as mediators between the data and the application domain worlds. In the sequel we are going to combine the three approaches of this section. The meta-model corresponds to other meta-models such as inductive modelling or hypothetical reasoning (hypotheses development, experimenting and testing, analysis of results, interim conclusions, reappraisal against real world).

4 Data Mining: A Systematic Model-Based Approach

Our approach presented so far allows to revise and to reformulate the model-oriented data mining process on the basis of well-defined engineering [15, 25] or alternatively on systematic mathematical problem solving [22]. Figure 4 displays this revision. We realize that the first two phases are typically implicitly assumed and not considered. We concentrate on the non-iterative form. Iterative processes can be handled in a similar form.

4.1 Setting the Deep Model and the Matrix

The problem to be tackled must be clearly stated in dependence on the utilization scenario, the tasks to be solved, the community of practice involved, and the given context. The result of this step is the deep model and its matrix. The first one is based on the background, the specific context parameter such as infrastructure and environment, and candidates for deep models.

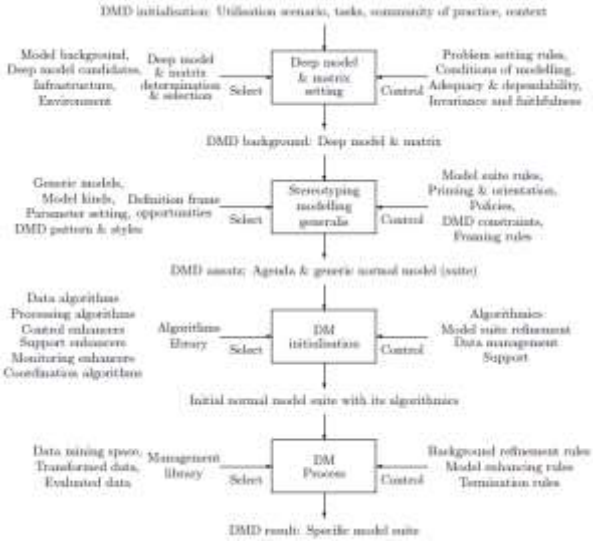


Figure 4 The Phases in Data Mining Design (Non-iterative form)

The data mining tasks can be now formulated based on the matrix and the deep model. We set up the context, the environment, the general goal of the problem and also criteria for *adequateness* and *dependability* of the solution, e.g. *invariance properties* for problem description and for the task setting and its mathematical formulation and *solution faithfulness properties* for later application of the solution in the given environment. What is exactly the problem, the expected benefit? What should a solution look like? What is known about the application?

Deep models already use a background consisting of an undisputable grounding and a selectable basis. The explicit statement of the background provides an understanding of the postulates, paradigms, assumptions, conceptions, practices, etc. Without the background, the results of the analysis cannot be properly understood. Models have their profile, i.e. goals, purposes and functions. These must be explicitly given. The parameters of a generic model can be either order or slave parameters [12], either primary or secondary or tertiary (also called genotypes or phenotypes or observables) [1, 5], and either ruling (or order) or driven parameters [12]. Data mining can be enhanced by knowledge management techniques.

Additionally, the concept space into which the data mining task is embedded must be specified. This concept space is enhanced during data analysis.

4.2 Stereotyping the Process

The general flow of data mining activities is typically implicitly assumed on the basis of stereotypes which form a set of tasks, e.g. tasks of prove in whatever system, transformation tasks, description tasks, and

investigation tasks. Proofs can follow the classical deductive or inductive setting. Also, abductive, adductive, hypothetical and other reasoning techniques are applicable. Stereotypes typically use model suites as a collection of associated models, are already biased by priming and orientation, follow policies, data mining design constraints, and framing.

Data mining and analysis is rather stereotyped. For instance, mathematical culture has already developed a good number of stereotypes for problem formulation. It is based on a mathematical language for the formulation of analysis tasks, on selection and instantiation of the best fitting variable space and the space of opportunities provided by mathematics.

Data mining uses *generic models* which are the basis of normal models. Models are based on a separation of concern according the problem setting: dependence-indicating, dependence-describing, sepa-ration or partition spaces, pattern kinds, reasoning kinds, etc. This separation of concern governs the classical data mining algorithmic classes: association analysis, cluster analysis, data grouping with or without classification, classifiers and rules, dependences among parameters and data subsets, predictor analysis, syner-getics, blind or informed or heuristic investigation of the search space, and pattern learning.

4.3 Initialization of the Normal Data Models

Data mining algorithms have their capacity and potential [2]. Potential and capacity can be based on SWOT (strengths, weaknesses, opportunities, and threats), SCOPE (situation, core competencies, obstacles, prospects, expectation), and SMART (how simple, meaningful, adequate, realistic, and trackable) analysis of methods and algorithms. Each of the algorithm classes has its strengths and weaknesses, its satisfaction of the tasks and the purpose, and its limits of applicability. Algorithm selection also includes an explicit specification of the order of application of these algorithms and of mapping parameters that are derived by means of one algorithm to those that are an input for the others, i.e. an explicit association within the model suite. Additionally, evaluation algorithms for the success criteria are selected. Algorithms have their own obstinacy, their hypotheses and assumptions that must be taken into consideration. Whether an algorithm can be considered depends on acceptance criteria derived in the previous two steps.

So, we ask: *What kind of model suite architecture suits the problem best? What are applicable development approaches for modelling? What is the best modelling technique to get the right model suite? What kind of reasoning is supported? What not? What are the limitations? Which pitfalls should be avoided?*

The result of the entire data mining process heavily depends on the appropriateness of the data sets, their properties and quality, and more generally the data schemata with essentially three components: application data schema with detailed description of data types, metadata schema [18], and generated and auxiliary data schemata. The first component is well-investigated in data mining and data management monographs. The second and third components inherit research results from database management, from data mart or warehouses, and layering of data. An essential element is the explicit specification of the quality of data. It allows to derive algorithms for data improvement and to derive limitations for applicability of algorithms. Auxiliary data support performance of the algorithms.

Therefore typical data-oriented questions are: *What data do we have available? Is the data relevant to the problem? Is it valid? Does it reflect our expectations? Is the data quality, quantity, recency sufficient? Which data we should concentrate on? How is the data transformed for modelling? How may we increase the quality of data?*

4.4 The Data Mining Process Itself

The data mining process can be understood as a coherent and stepwise refinement of the given model suite. The model refinement may use an explicit transformation or an extract-transform-load process among models within the model suite. Evaluation and termination algorithms are an essential element of any data mining algorithm. They can be based on quality criteria for the finalized models in the model suite, e.g. generality, error-proneness, stability, selection-proneness, validation, understandability, repeatability, usability, usefulness, and novelty.

Typical questions to answer within this process are: *How good is the model suite in terms of the task setting? What have we really learned about the application domain? What is the real adequacy and dependability of the models in the model suite? How these models can be deployed best? How do we know that the models in the model suite are still valid? Which data are supporting which model in the model suite? Which kind of errors of data is inherited by which part of which model?*

The final result of the data mining process is then a combination of the deep model and the normal model whereas the first one is a latent or hidden component in most cases. If we want, however, to reason on the results then the deep model must be understood as well. Otherwise, the results may become surprising and may not be convincing.

4.5 Controllers and Selectors

Algorithmics [6] treats algorithms as general solution pattern that have parameters for their instantiation, handling mechanisms for their specialization to a given environment, and enhancers for context injection. So, an algorithm can be derived based on explicit selectors and control rules [4] if we neglect context injection. We can use this approach for data mining design (DMD).For instance, an algorithm pattern such as regression uses a generic model of parameter dependence, is based on blind search, has parameters for similarity and model quality, and has selection support for specific treatment of the given data set. In this case, the controller is based on enablers that specify applicability of the approach, on error rules, on data evaluation rules that detect

dependencies among control parameters and derive data quality measures, and on quality rules for confidence statements.

4.7 Data Mining and Design Science

Let us finally associate our approach with design science research [13]. Design science considers systematic modelling as an embodiment of three closely related cycles of activities. The *relevance cycle* initiates design science research with an application context that not only provides the requirements for the research as inputs but also defines acceptance criteria for the ultimate evaluation of the research results. The central *design cycle* iterates between the core activities of building and evaluating the design artifacts and processes of the research. The orthogonal *rigor cycle* provides past knowledge to the research project to ensure its innovation. It is contingent on the researchers' thoroughly research and references the knowledge base in order to guarantee that the designs produced are research contributions and not routine designs based upon the application of well-known processes.

The relevance cycle is concerned with the problem specification and setting and the matrix and agenda derivation. The design cycle is related to all other phases of our framework. The rigor cycle is enhanced by our framework and provides thus a systematic modelling approach.

5 Conclusion

The literature on data mining is fairly rich. Mining tools have already gained the maturity for supporting any kind of data analysis if the data mining problem is well understood, the intentions for models are properly understood, and if the problem is professionally set up. Data mining aims at development of model suites that allows to derive and to draw dependable and thus justifiable conclusions on the given data set. Data mining is a process that can be based on a framework for systematic modelling that is driven by a deep model and a matrix. Textbooks on data mining typically explore in detail algorithms as blind search. Data mining is a specific form of modeling. Therefore, we can combine modeling with data mining in a more sophisticated form. Models have however an inner structure with parts which are given by the application, by the context, by the commonsense and by a community of practice. These fixed parts are then enhanced by normal models. A typical normal model is the result of a data mining process.

The current state of the art in data mining is mainly technology and algorithm driven. The problem selection is made on intuition and experience. So, the matrix and the deep model are latent and hidden. The problem specification is not explicit. Therefore, this paper aims at the entire data mining process and highlights a way to leave the ad-hoc, blind and somehow chaotic data analysis. The approach we are developing integrates the theory of models, the theory of problem solving, design science, and knowledge and content management. We realized that data mining can be systematized. The framework for data mining design exemplarily presented is an example in Figure 4.

Acknowledgement

We thank for the support of this paper by the CRC 1266. We are very thankful for the fruitful discussions with the members of the CRC.

References

[1] Bell, G.: The mechanism of evolution. Chapmanand Hall, NewYork (1997)

[2] Berghammer, R., Thalheim, B.: Metho-denbasiertemathematische Modellierungmit Relationenalgebren. In: Wissenschaft und Kunst der Modellierung: Modelle, Modellieren, Modellierung, pp. 67–106. DeGryuter, Boston (2015)

[3] Berthold, M.R., Borgelt, C., H¨oppner, F., Klawonn, F.: Guidetointelligentdata analysis. Springer, London (2010)

[4] Bienemann, A., Schewe, K.-D., Thalheim, B.: Towards a Theory of Genericity Based on Government and Binding. In: Proc. ER'06, LNCS, 4215, pp. 311-324. Springer (2006)

[5] Booker, L.B., Goldberg, D.E., Holland, J.H.: Classifier Systems and Genetic Algorithms. Artificial Intelligence, 40 (1–3): pp. 235-282 (1989)

[6] Brassard, G., Bratley, P.: Algorithmics – Theory and Practice. Prentice Hall, London (1988)

[7] Coleman, A. Scientific Models as Works. Cataloging&Classification Quarterly, Special Issue: Works as Entities for Information Retrieval, 33, pp. 3-4 (2006)

[8] Dahanayake, A., Thalheim, B.: Co-evolution of (information) System Models. In: EMMSAD2010, LNBIB, 50, pp. 314-326. Springer (2010)

[9] Embley D., Thalheim B. (eds): The Handbook of Conceptual Modeling: Its Usage and Its Challenges. Springer (2011)

[10] Gillett, N.P., Zwiers, F.W., Weaver, A.J., Hegerl, G.C., Allen, M.R., Stott, P.A.: Detecting Anthropogenic Influence with Amulti-modelensemble. Geophys. Res. Lett., 29, pp. 31-34 (2002)

[11] Guerra, E., deLara, J., Kolovos, D.S., Paige, R.F.: Inter-modelling: From Theory to Practice. In MoDELS2010, LNCS, 6394, pp. 376-391, Springer (2010)

[12] Haken, H., Wunderlin, A., Yigitbasi, S.: An Introduction to Synergetics. Open Systems and Information Dynamics, 3 (1), pp. 1-34 (1994)

[13] Hevner, A. March, S., Park, J., Ram, S.: Design Science in Information Systems Research. MIS Quaterly, 28 (1), pp. 75-105 (2004)

[14] Hunter, P.J., Li, W.W., McCulloch, A.D., Noble, D.: Multiscale Modeling: Physiome Project

Standards, Tools, and Databases. IEEEComputer, 39 (11), pp. 48-54 (2006)

[15] ISO/IEC 25020 (Software and Systemen Gineering – Software Product Quality Requirements and Evaluation (square) – Measurement Reference Model and Guide). ISO/IEC JTC1/SC7N3280 (2005)

[16] Jaakkola, H., Thalheim, B., Kidawara, Y., Zettsu, K., Chen, Y., Heimbu¨rger, A.: Information Modeling and Global Risk Management Systems. In: Information Modelingand Knowledge Bases, XX, pp. 429-446. IOS Press (2009)

[17] Jannaschk, K.: Infrastruktur für ein Data Mining Design Framework. PhDthesis, Christian-Albrechts University, Kiel (2017)

[18] Kramer, F., Thalheim, B.: A Metadata System for Quality Management. In: Information Modelling and Knowledge Bases, pp. 224-242. IOS Press (2014)

[19] Nakoinz, O., Knitter, D.: Modelling Human Behaviour in Landscapes. Springer (2016)

[20] Pardillo, J.: A Systematic Review on the Definition of UML Profiles. In: MoDELS2010, LNCS, 6394, pp. 407-422, Springer (2010)

[21] Petrelli D., Levin, S., Beaulieu, M., Sanderson, M.: Which User Interaction for Cross-language Information Retrieval? Design Issues and Reflections. JASIST, 57 (5), pp. 709-722 (2006)

[22] Pilkeyand, O.H., Pilkey-Jarvis, L.: Useless Arithmetic: Why Environmental Scientists Cant't Predict the Future. Columbia University Press, NewYork (2006)

[23] Podkolsin, A. S.: Computer-based Modelling of Solution Processes for Mathematical Tasks (in Russian). ZPI at Mech-Mat MGU, Moscov (2001)

[24] Pottmann, M., Unbehauen, H., Seborg, D.E.: Application of Ageneral Multi-model Approach for Identification of Highly Nonlinear Processes – a Case Study. Int. J. of Control, 57 (1), pp. 97-120 (1993)

[25] Rumpe, B.: Modellierungmit UML. Springer, Heidelberg (2012)

[26] Samuel, A., Weir, J.: Introduction to Engineering: Modelling, Synthesis and Problem Solving Strategies. Elsevier, Amsterdam (2000)

[27] Simsion, G., Witt, G.C.: Data Modeling Essentials. Morgan Kaufmann, San Francisco (2005)

[28] Skusa, M.: Semantische Kohärenz inder Softwareentwicklung. PhDthesis, CAU Kiel, (2011)

[29] Thalheim, B.: Towards a Theory of Conceptual Modelling. J. of Universal Computer Science, 16 (20), pp. 3102-3137 (2010)

[30] Thalheim, B.: The conceptual model ≡ an Adequate and Dependable Artifacten Hanced Byconcepts. In: Information Modellingand Knowledge Bases XXV, pp. 241-254. IOS Press (2014)

[31] Thalheim, B.: Conceptual Modelling Foundations: The Notion of a Model in Conceptual Modeling. In: Encyclopedia of Database Systems, Springer (2017)

[32] Thalheim, B., Tropmann-Frick, M.: Where Foremodels are Used and Accepted? The Model Functions as Aquality Instrument Inutilisation Scenarios. In: I. Comyn-Wattiau, C. duMouza, N. Prat, editors, Ingenierie Management des Systemes D'Information (2016)

[33] Thalheim, B., Tropmann-Frick, M., Ziebermayr, T.: Application of Generic Workflows for Disaster Management. In: Information Modellingand Knowledge Bases, pp. 64-81. IOS Press (2014)

[34] Thalheim, B., Wang, Q.: Towards a Theory of Refinement for Data Migration. In: ER'2011, LNCS, 6998, pp. 318-331. Springer (2011)

[35] Zeugmann, T.: Inductive Inference of Optimal Programs: A Survey and Open Problems. In: Nonmonotonic and Inductive Logics, pp. 208-222. Springer (1991)

Оценка качества научных гипотез в виртуальных экспериментах в областях с интенсивным использованием данных

© Е.А. Тарасов[1] © Д.Ю. Ковалев[2]

[1]Московский государственный университет имени М.В. Ломоносова,
[2]Федеральный исследовательский центр «Информатика и управление» РАН,
Москва, Россия

tarasov@outlook.com dkovalev@ipiran.rue

Аннотация. Исследованы подходы, позволяющие оценить качество модели, реализующей гипотезу в рамках виртуального эксперимента. Математические модели, порождающие гипотезы, активно используются в областях с интенсивным использованием данных. К одной из таких областей можно отнести исследования многофазных потоков жидкости. Модель реализует гипотезу о скорости потока жидкости в трубе. Подход оценки качества осуществляется в рамках общей классической теории детектирования сигнала. Оценка представляет собой бинарный показатель. В качестве аппарата проверки гипотез используются два метода: частотный и Байесовский. Обработка входных данных осуществляется в потоковом режиме. Из данных, поступающих на вход системе, на этапе предобработки извлекаются признаки, которые затем подаются на вход исследуемой модели. С определенной периодичностью происходит перерасчет оценки качества с учетом изменяющихся параметров среды. Таким образом, отслеживается момент, когда модель начинает плохо предсказывать поведение физического явления. В работе описана реализация данного функционала на распределенной вычислительной системе.

Ключевые слова: виртуальный эксперимент, управление гипотезами, частотный подход, Байесовский подход, многофазное течение жидкости.

Estimation of Scientific Hypotheses Quality in Virtual Experiments in Data Intensive Domains

© Evgeny Tarasov[1] © Dmitry Kovalev[2]

[1] Lomonosov Moscow State University,
[2] Federal Research Center Computer Science and Control of the Russian Academy of Sciences,
Moscow, Russia

tarasov@outlook.com dkovalev@ipiran.rue

Abstract. In this paper, we investigate approaches that allow us to estimate the quality of model implementing hypotheses within a virtual experiment. One of the areas of DID under study is the multiphase fluid flow analyses. The quality estimation approach is carried out within the framework of the general classical detection theory. The estimate is a binary indicator. As a instrument for testing hypotheses, two approaches are used: frequency and Bayesian. Feature Extraction is carried out in the streaming mode. With a certain periodicity, the quality assessment is recomputed taking into account the changing environmental parameters. Thus, the moment when the model begins to poorly predict the behavior of the physical phenomenon is captured. This paper describes the implementation of this approach within a distributed computing framework.

Keywords: virtual experiment, hypothesis management, frequency approach, Bayesian approach, multiphase fluid flow.

1 Введение

Данные в современных исследованиях имеют определяющую роль [11]. Они могут быть представлены как в неструктурированном, так и в полуструктурированном виде. Таким образом, акцент в работе ученого смещается с проведения реального физического эксперимента на обработку данных в рамках виртуального эксперимента, моделирующего поведение физического явления. Данная парадигма работы получила название исследований с интенсивным использованием данных (ИИИД) [20].

Труды XIX Международной конференции «Аналитика и управление данными в областях с интенсивным использованием данных» (DAMDID/ RCDL'2017), Москва, Россия, 10–13 октября 2017 года

К одной из областей ИИИД можно отнести течение многофазных потоков жидкости [8]. Данные собираются с множества сенсоров, например, выполняется дискретная запись акустического сигнала, температур, давления, течения жидкости. Накопленные объемы данных служат для характеристики многофазных потоков и их режимов. Важным приложением интерпретации акустического сигнала и накопленных мета-данных является предсказание скорости потока жидкости.

Несмотря на значительные успехи в интерпретации данных с сенсоров, проблемы построение сложных моделей, объясняющих динамику течения жидкости, остаются открытыми. Проведение точного физического эксперимента требует от исследователя серьезных усилий, т. к. необходимо обеспечить множество специальных условий [32]. Численное моделирование, особенно для недостаточно хорошо изученных потоков, часто требует калибровки с экспериментом [25, 31]. Разработка подходов к анализу данных для интерпретации потоковых данных виртуального эксперимента является важной и перспективной проблемой для исследования.

Параметры модели течения жидкости не являются статичным элементом. Внешние условия изменчивы и влияют на неё. Из-за этого предсказательная способность модели может начинать резко ухудшаться [30]. Это приводит к необходимости проведения её повторной калибровки. Важно отследить тот момент, когда модель начинает выдавать заведомо плохой результат. Таким образом, платформа, обрабатывающая вычисления, должна уметь работать с потоковыми данными в режиме, близком к реальному времени. Это обуславливает особое отношение к такому роду задач.

Одной из таких особенностей является требование распределенности. Система должна быть легко масштабируема, чтобы обрабатывать модель любой сложности с постоянно возрастающим объемом данных за разумный промежуток времени. Готовых открытых систем в области гидродинамики, удовлетворяющих данному требованию, по сведению авторов, нет.

Всё больше задач формируется не в рамках одной области, а междисциплинарно. Таким образом, возрастает роль онтологических спецификаций [37]. Это позволяет как различным ученым в рамках одной области, так и ученым из различных областей использовать общие понятия, что необходимо для ускорения проведения совместных исследований. В данной работе область применения исследования лежит как в контексте изучения физических явлений, а именно, течения жидкости, так и управления гипотезами [10].

Новизной данной работы является распределенная реализация метода бинарной оценки качества модели, работающей с потоковыми данными.

Статья организована следующим образом. В разделе 2 представлен сравнительный обзор платформ, на базе которых может быть осуществлен виртуальный эксперимент. Раздел 3 определяет онтологическую спецификацию двух предметных областей: течения жидкости и управления гипотезами. В разделе 4 выполнена концептуальная спецификация данных, используемых в виртуальном эксперименте. Раздел 5 описывает формат сырых входных данных, данных, поступающих на вход модели, и извлечённые значимые признаки модели течения жидкости. Раздел 6 содержит описание метода оценки качества с использованием двух подходов: частотного и Байесовского. Раздел 7 раскрывает архитектуру использованного для расчётов программно-аппаратного комплекса. В разделе 8 представлен реализованный поток работ. Раздел 9 описывает результаты, полученные на практике.

2 Обзор платформ

2.1 Критерии выбора

Одним из ключевых элементов ИИИД, наряду с машинным обучением, является явное использование гипотез в определении виртуального эксперимента [10]. Многие исследователи скептически относятся к подходу с использованием машинного обучения, так как он дает низкую интерпретируемость полученных результатов, так как многие методы в нем используются как черный ящик [18]. Подход же на основе гипотез лишен данного недостатка. Гипотезы в математическом виде описывают априорные знания об исследуемом явлении, которые проверяются в рамках виртуального эксперимента.

На сегодняшний день отсутствует единая методология работы с потоковыми данными [36]. Программные продукты, которые в той или иной мере реализуют представления отдельных групп ученых на то, как нужно работать с ними, не являются развитыми и стабильными.

В рамках исследования существующих решений по поставленной перед нами цели рассмотрим платформы для обработки данных в таких области научной деятельности как:

- средства управления гипотезами и проведения виртуального эксперимента;
- средства управления и обработки потоковых данных.

В качестве требований, предъявляемым при сравнительном анализе систем, будем применять следующие положения:

1. Система должна соответствовать онтологической спецификации предметной области. Это значит, что на базе нее возможно:
   a. реализовать модель;
   b. провести статистическое тестирование;
5. Система должна уметь работать с большим объемом потоковых данных;

Распределенность и скорость, т. е. должна легко масштабироваться в зависимости от вычислительной нагрузки. Результат должен получаться в режиме, близком к реальному времени;

Таблица 1 Сравнение платформ управления гипотезами

Название	Ключевые элементы	Слабые стороны	Сильные стороны
Hephaestus	Собственный SQL-подобный язык запросов для описания эксперимента. Основной элемент – виртуальный эксперимент. Построение вероятностно-причинных графов. Мета-система – работает над существующими базами данных.	Интеграция данных не предоставляется из коробки. Плохое описание модуля поиска корреляций. Ориентированность применения – здравоохранение, ведет к своей интерпретации определения виртуального эксперимента.	Тестирование и ранжирование гипотез на основе частотной статистики. Граф знаний может визуализировать результаты и помогать интегрировать новые гипотезы. Работает с наборами гипотез.
FCCE	Использование NoSQL БД. Основной элемент – концепция функций. API поддержки для хранения, извлечения, оценки корреляции по признакам. Комплексная многоуровневая система агрегации данных.	Не описан модуль корреляций. Ориентированность применения – анализ поведения сети. Не оперирует математическими формулами.	Сосредоточение на минимизации задержек. Поддержка доступа к сырым данным. Быстрый модуль поиска корреляций. Программно реализован.
ϒ-DB	Основной элемент – поддержка научных исследований. Вероятностная БД. Работа с отдельными гипотезами. Байесовский подход. Гипотезы в формате MathML.	Взаимосвязь гипотез выходит за рамки системы. Проблемы с масштабируемостью.	СУБД на базе SQL. Автоматически пересчитывает вероятность после получения новых данных или гипотезы. Работает с формулами.

6. Открытость, т. е. система должны быть с открытым исходным кодом;
7. Стабильность работы также является важным критерием, т. к. многие средства были разработаны ещё совсем недавно и не прошли полного цикла отладки.

2.2 Управление гипотезами

В настоящее время в рамках работы в областях с интенсивным использованием данных многие исследователи приходят к выводу о необходимости унификации подходов построения виртуальных экспериментов. Отдельные научные группы разрабатывают свои программные продукты, реализующие видение своих авторов к данной проблематике. Проанализируем некоторые из них – такие продукты, как: Hephaestus [10, 38], Features Collection and Correlation Engine (FCCE) [26, 38], ϒ-DB [11, 12, 38]. Сводная информация по сравнению платформ управления гипотезами представлена в Таблице 1.

2.3 Потоковые системы

Для сравнения выберем открытые системы, которые являются наиболее популярными с точки зрения применимости в практических задачах. К таким продуктам можно отнести: Storm [5, 21, 15], Flink [1, 2, 15], Spark Streaming [4, 15, 27, 13].

Серьёзной проблемой выбора потоковых фреймворков [36, 29] является отсутствие в настоящее время единых и объективных критериев оценки производительности [36]. В научной литературе существуют публикации, авторы которых проводят определённые сравнения, однако проблемой являются узкая специализация и ограниченность применения этих тестов [9]. В связи с отсутствием единой методологии целесообразно получать сравнительную производительность систем на данных конкретной исследовательской задачи для всех анализируемых платформ.

Однако скорость является не единственным критерием выбора платформы. Сводные данные [13, 15] по потоковым системам представлены в Таблице 2.

2.4 Выбор платформы

Исходя из всех выше приведенных обзоров, можно сделать следующие выводы.

В рамках данной работы невозможно использовать никакую из существующих систем по управлению гипотезами. Это обусловлено тем, что они:

- ориентированы на работу с статическими данными, хранящимися в базе данных;
- нет гибкого инструмента построение виртуального эксперимента; имеется ориентаций на свои области применения;
- не являются открытыми программными продуктами;
- не носят законченного характера; некоторые модули имеют только описательный характер без практической реализации;
- плохая документированность.

Таблица 2 Сравнение потоковых систем

Критерий	**Storm**	**Flink**	**Spark Streaming**
Поддержка языков программирования	Java	Java, Scala	Java, Scala, Python
Режим работы	Потоковый (tuple-wise)	Потоковый и микро-пакеты	Микро-пакеты (micro-batch)
Обработка сообщений	По крайней мере один раз (at least once)	Строго один раз (exactly once)	Строго один раз (exactly once)
Управление окном	Нет, встроенными средствами	На основе: времени, строк, приходящих данных	Только на основе времени
Управление ресурсами	YARN, Mesos, Built-in	YARN, Built-in	YARN, Mesos, Built-in
Задержка обработки	Низкая	Низкая	Средняя
Механизмы обеспечения отказоустойчивости	ACK записи	Распределенные снапшоты	Микро-пакет
Управление потоком	Проблематично	Естественно	Проблематично
Операции с сохранением состояния	Нет	Да	Да
Потоковая примитива	Tuple	DataStream	DStream
Пропускная способность	Низкая	Высокая	Высокая

В качестве платформы оценки качества виртуального эксперимента выбрана платформа на базе Spark Streaming. Этот выбор обусловлен следующими положениями:

- текущая реализация модели течения жидкости представлена на Python;
- имеется возможность расширения функционала программирования проверки статистических гипотез за счет установки дополнительных библиотек;
- возможность управления окном на основе времени.

Из сравнительной таблицы потоковых систем видно, что Spark Streaming не является самой производительной системой. Её выбор обусловлен поддержкой языка Python и обеспечением минимально необходимого функционала.

3 Онтологическая спецификация

3.1 Течение жидкости

Онтологическая спецификация течения жидкости представлена на Рис. 1. Данная онтология описывает проведение реального эксперимента, при котором анализируется течение жидкости в трубе [8]. В качестве жидкости могут использоваться: вода, масло, их смесь с газом. Жидкость в рамках данного эксперимента обладает следующими свойствами: температура, скорость потока, давление. В рамках модели акустический шум и спектр являются функциями давления. Спектр в свою очередь характеризуется амплитудой и частотой, которые являются входными данными для определения величины частотного пика. Давление измеряется гидрофонами, установленными в гидродинамической трубе. Также для измерения температуры скорости потока используются дополнительные сенсоры. Данные с гидрофонов и сенсоров поступают на аналогово-цифровой преобразователь, который, в свою очередь, усиливает и дискретизирует поступающий сигнал в соответствии с заданными свойствами. В эксперименте частота дискретизации составляет 100 кГц. Обработанный сигнал записывается в результирующий файл эксперимента.

Онтология предметной области соответствует реальному физическому эксперименту. Каждый эксперимент выполняется в течение 25 секунд. Затем происходит некоторое изменение в параметрах модели, затем эксперимент повторяется снова. Так на выходе получается набор результирующих файлов. Поток данных формируется искусственно с использованием этих файлов.

3.2 Управление гипотезами

Онтологическая спецификация управления гипотезами представлена на Рис. 2.

Определим некоторые термины предметной области. Виртуальный эксперимент – деятельность по применению набора гипотез для воспроизведения симуляций близких к наблюдаемому явлению. Гипотеза – формальная спецификация свойств исследуемого объекта или явления, имеющих математическое представление. Модель – алгоритм, реализующий гипотезы. Оценка качества – характеристика модели, позволяющая сделать вывод о её соответствии реальному явлению.

Статистическое тестирование – это процесс принятия решения о том, противоречит ли рассматриваемая статистическая гипотеза выборке данных. Частотный подход – аппарат проверки гипотез, базирующийся на частотном определении вероятности, т. е. вероятности как предела относительной частоты наблюдения некоторого события в серии однородных независимых испытаний. Байесовский подход – аппарат проверки гипотез, базирующийся на байесовском определении вероятности, для которой имеются некоторые априорные знания о наблюдаемом явлении. Потоковая система – программное средство, позволяющее анализировать непрерывно поступающие на ее вход данные.

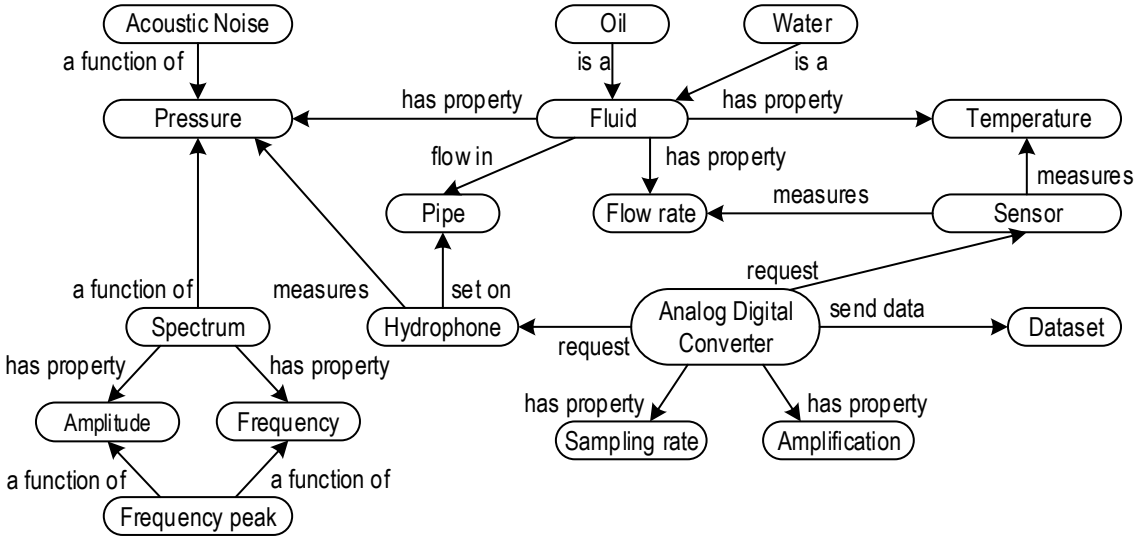


Рисунок 1 Онтология течения жидкости

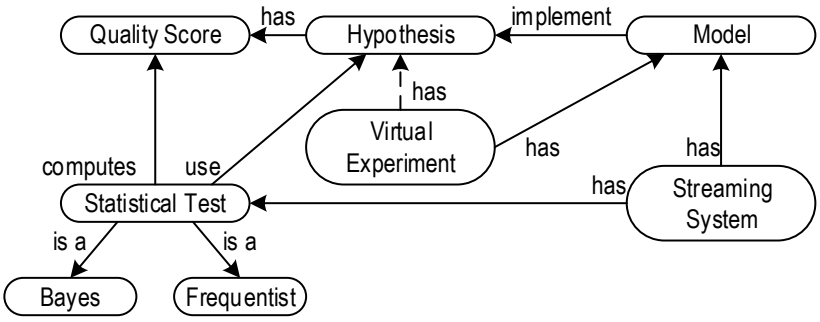


Рисунок 2 Онтология управления гипотезами

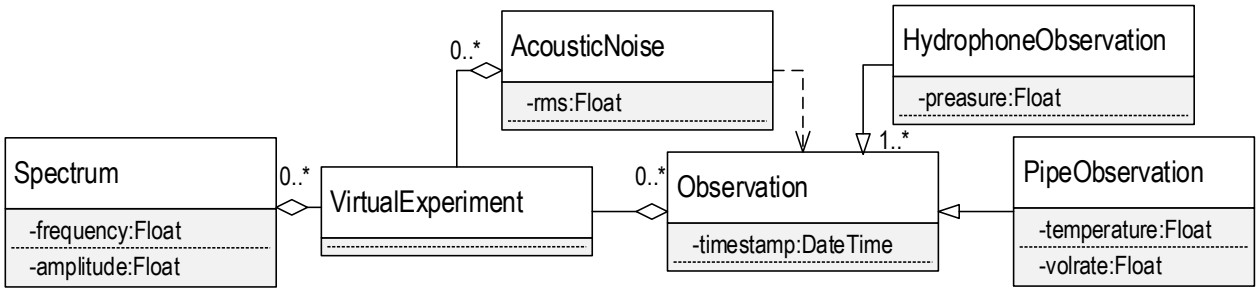


Рисунок 3 Концептуальная спецификация

На базе потоковой системы реализована модель, описывающая течение жидкости в трубе. Также данная система имеет модули выполнения проверки статистических тестов. На базе результата работы модели и входных данных сформулированы гипотезы, позволяющие получить оценку качества модели. Оценка высчитывается на базе статистических тестов, заложенных в потоковой системе. В виртуальном эксперименте рассматриваются частотный и Байесовский подходы для получения оценки качества.

4 Концептуальная спецификация

Концептуальная спецификация виртуального эксперимента представлена на Рис. 3.

В рамках исследуемой модели имеется априорное знание о зависимости скорости потока от температуры, значении частоты пика спектра, шума потока. Также известно, что имеются корреляции между показателями гидрофонов, установленных в гидродинамической трубе. Данные разделяются на полученные из наблюдения и вычисленные из данных наблюдения. К непосредственно получаемым данным относится информация от гидрофонов и сенсоров, а именно: давление, температура, скорость потока. К вычисляемым относятся: спектр, амплитуда, акустический шум. Все эти характеристики рассчитываются на основе наблюдаемых показателей давления. Все данные, кроме временной метки, имеют вещественный тип.

5 Формат данных

5.1 Входные данные

Данные, полученные в рамках реального

физического эксперимента, хранятся в CSV (Comma Separated Value) файлах на распределенной файловой системе (HDFS). На вход системе поступает два вида файлов: данные, полученные с гидрофонов; данные, полученные с сенсоров.

Файл с информацией, полученной с гидрофонов, имеет следующее описание. Заголовок файла состоит из порядкового номера *nm*,– указатель канала гидрофона $h\{i\}$, где i – число от 0 до 2. Столбцы разделены знаком табуляции. Заголовок файла выглядит так: `nm\t\h0\h1\h2`.

Частота дискретизации для сбора акустических данных составляет 100 кГц. Один эксперимент длится 25 секунд. Таким образом, каждый CSV файл содержит 2,5 миллиона строк.

Файл с информацией, полученной с сенсоров, имеет следующее описание: заголовок файла состоит из timestamp – временной метки, temp – температуры, vol_rate – скорости потока, file – файла, соответствующего показателям гидрофонов. Столбцы разделены знаком ";". Заголовок файла выглядит так: `timestamp;temp;vol_rate;file`

Показатели температуры и скорости потока собираются один раз в секунду. Вопросы синхронизации показателей различных файлов решаются через информацию временных меток и сопоставления имен файлов в соответствующих полях.

5.2 Обрабатываемые данные

Данные для обработки в потоковой системе преобразуются в формат RDD (Resilient Distributed Dataset) – отказоустойчивый набор элементов, обработка которых может выполняться параллельно [28]. Вся логика работы с данными происходит в рамках этого концепта. Существуют два способа получения такого набора:

- параллелизация последовательного массива в рамках текущей программы с помощью вызова метода parallelize();
- на этапе извлечения данных из внешних источников, таких, как HDFS, HBase, Streaming Context.

При создании распределенного набора можно как явно задать число, показывающее, сколько параллельных разделов использовать при работе с этими данными, так и использовать значение по умолчанию.

В рамках виртуального эксперимента самые «тяжелые» в вычислительном плане задачи ложатся на набор данных, поступивших с гидрофонов. Так как используется информация из трёх каналов, то степень параллелизма на этапе предобработки устанавливается также равной трём.

Концепт RDD имеет свой API и поддерживает два вида операций:

- трансформацию – получение нового набора из существующего;
- действие – запускает задание на выполнение.

При написании программы важно отслеживать область действия переменных в рамках данных видов операций.

Важными характеристиками RDD являются:

- распределенность – операции над данными выполняются на различных узлах кластера;
- ленивое исполнение ("lazy") – трансформация не выполняется прямо сейчас; система хранит последовательность операций над набором; выполнение происходит, только если в коде программы встретилась операция действия;
- управление состоянием – возможность выбора, из какого хранилища (память или диск) они будут повторно использоваться.

6 Подход оценки качества

6.1 Описание метода

Рассматриваемый нами метод оценки качества виртуального эксперимента опирается на классическую теорию детектирования сигнала [14]. Эта теория выступает как средство количественной оценки возможности различать информационную составляющую сигнала от шума [24]. Базовые компоненты теории детектирования представлены на Рис. 4.

Первым ее элементом является источник, который генерируют выходной сигнал. Его выход может быть одним из нескольких вариантов. В самом простом случае это гипотезы: H_1 и H_0. Второй и третий компоненты соответственно: механизм вероятностного перехода и пространство наблюдений. Механизм перехода может рассматриваться для определения, какая гипотеза истина. На основе этих знаний он генерирует точку в пространстве наблюдений в соответствии с некоторым законом вероятности. Независимая дискретная случайная величина n, чья плотность вероятности нам известна, добавляется к выходу источника.

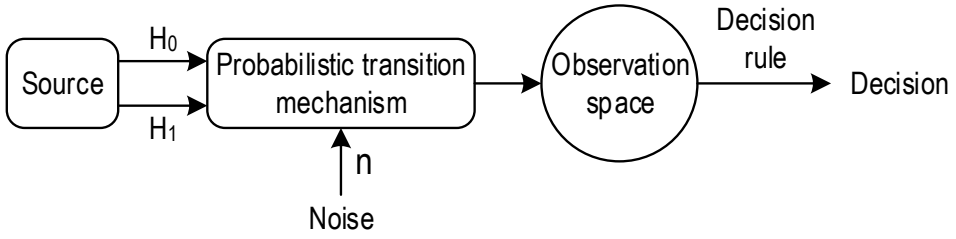


Рисунок 4 Компоненты теории детектирования

Четвертым компонентом теории детектирования сигнала является правило принятия решения. После получения наблюдаемой переменой в пространстве наблюдений мы должны угадать какая гипотеза была истинной. Правило принятия решений сопоставляет каждой точке наблюдений одну из гипотезПодходящий выбор правил будет зависеть от многих факторов, которые определяются постановкой конкретной исследовательской задачи.

Мы рассматриваем оценку качества как бинарное событие, которое в математическом представлении определено в виде гипотез:

H_0: модель корректна;

H_1: происходит нарушение работы модели.

Чтобы сделать предположение о качестве модели, необходимо выполнить статистическую проверку данных.

Одним из параметров, передаваемых на вход нашей системы один раз в секунду, является измеренная величина Y скорости потока. В рамках проведения виртуального эксперимента модель выдавала вычисленную величину $\hat{Y}$ скорости потока. Оценивая их разность $\hat{Y} - Y$ на протяжении 25 секунд, можно сделать вывод об оценке качества модели при заданных условиях эксперимента.

Известно, что разность $\hat{Y} - Y$ подчиняется нормальному закону распределения [8]. Таким образом, в качестве аппарата проверки гипотез могут быть использованы следующие подходы: классический или частотный метод; Байесовский метод.

6.2 Частотный подход

В рамках классического подхода вероятность определяется как относительная частота наступления события. Все события имеют независимый характер. Общая методика проведения статистического тестирования широко представлена в литературе (см., например, [35]). В рамках настоящей работы решалась задача в следующей постановке.

В исследуемой модели течения жидкости используем следующие положения:

- остатки подчиняются нормальному закону распределения;
- дисперсия сигнала неизвестна.

Таким образом, в рамках частотного подходы будет использован одно-выборочный Т-тест с двухсторонней альтернативой.

В качестве исследуемой случайной величины возьмем разность $X_n = \hat{Y} - Y$ контрольного значения скорости потока, поступающего с датчиков, и величины скорости потока, полученной в результате работы модели. Таким образом, получим выборку

$$X = (x_1, \dots, x_n) \in R, \quad X_n \sim N(\mu, \sigma^2).$$

Проверим гипотезу, что выборочное среднее равно заданному числу, против альтернативной гипотезы, что это не так: $H_0: \bar{X} = \mu$, $H_1: \bar{X} \neq \mu$. В нашем случае примем $\mu = 0$. В качестве правила принятия решения возьмем критерий Стьюдента [34]. Статистика критерия имеет распределение Стьюдента с $n-1$ степенями свободы:

$$T(X) = \frac{(\bar{X} - \mu)\sqrt{n}}{S} \sim st(n-1),$$

где выборочное среднее $\bar{X} = \frac{1}{n}\sum_{i=1}^{n} x_i$, выборочная дисперсия $S^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{X})^2$.

Пусть задан уровень значимости $\alpha = 0.05$. В нашей задаче используем двухстороннюю альтернативу. Таким образом, статистический критерий будет иметь вид $|T| > T_{\alpha/2}$, где T_α – α-квантиль распределения Стьюдента с $n-1$ степенями свободы. Если $|T| > T_{\alpha/2}$, то нулевая гипотеза H_0 отвергается.

6.3 Байесовский подход

Байесовский подход отличается от классического тем, что в своей основе имеет другое определение вероятности (она интерпретируется как мера незнания, а не как объективная случайность [33]). В общем виде вероятность определяется как степень уверенности в истинности суждения. Мы имеем некоторое априорное знание о наблюдении, которое уточняется в процессе эксперимента.

Байесовская проверка в рамках теории детектирования базируется на двух предположениях:

- выходы источника регулируются вероятностным присвоением, они обозначаются P_1 и P_0 и называются априорными вероятностями, которые представляют собой информацию об источнике до проведения эксперимента;
- каждому из возможных исходов присваивается стоимость; обозначим стоимости 4-х исходов C_{00}, C_{10}, C_{01}, C_{11}; первый индекс указывает на выбранную гипотезу, второй – на ту, которая истинна; после каждого эксперимента стоимости могут уточняться.
- Правило принятия решения должно быть таким, чтобы средняя стоимость была как можно меньше. В общем виде данный подход характеризуется формулой [14]:

$$\frac{p_{r|H_1}(R|H_1)}{p_{r|H_0}(R|H_0)} \underset{H_0}{\overset{H_1}{\gtrless}} \frac{P_0(C_{10}-C_{00})}{P_1(C_{01}-C_{11})}.$$

Величина в левой части называется коэффициентом правдоподобия и обозначается $\Lambda(R)$. Величина в правой части формулы является пороговым значением теста и обозначается η.

Таким образом, Байесовский критерий приводит нас к проверке неравенств

$$\Lambda(R) \underset{H_0}{\overset{H_1}{\gtrless}} \eta \quad \text{или} \quad \ln\Lambda(R) \underset{H_0}{\overset{H_1}{\gtrless}} \ln\eta.$$

Эмпирическая шкала [20] доказательной силы Байесовского критерия приведена в Таблице 3.

Таблица 3 Шкала критерия правдоподобия

$\lvert \ln \eta \rvert$	η	Доказательная сила
< 1.0	< 3:1	Не убедительная
1.0	~ 3:1	Слабое доказательство
2.5	~ 12:1	Среднее доказательство
5.0	~ 150:1	Сильное доказательство

Доказательная сила говорит о том, можем мы или нет отвергнуть гипотезу H_0. Так как гипотеза H_0 предполагает, что наша модель корректна, то ее отвержение служит оценкой качества и указывает на то, что модель начинает плохо описывать наблюдения.

Дадим постановку задачи. В качестве исследуемой случайной величины, как и в частотном подходе, используем $X_n = \hat{Y} - Y$ (разность контрольного значения скорости потока, поступающего с датчиков, и величины скорости потока, полученной в результате работы модели). Таким образом, задана выборка

$$X = (x_1, \dots, x_n) \in R, X_n \sim N(\mu, \sigma^2).$$

Проверим гипотезу о том, что разность средних наблюдаемого и моделируемого сигналов равна нулю плюс имеется дополнительная составляющая некоторого шума, против альтернативной гипотезы о том, что кроме шума имеется ненулевая составляющая, информирующая, что сигнал ушел с базовой линии:

$$H_0: r_i = 0 + n_i = n_i, \ \ i = 1,2, \dots, N,$$
$$H_1: r_i = \mu + n_i, i = 1,2, \dots, N.$$

Шумовая составляющая подчиняется нормальному закону распределения. Таким образом,

$$p_i(X) = \frac{1}{\sqrt{2\pi}\sigma} \exp\left(-\frac{X^2}{2\sigma^2}\right).$$

Плотность вероятностей r_i при каждой гипотезе вычисляется следующим образом:

$$p_{r_i|H_1}(R_i|H_1) = p_{n_i}(R_i - \mu) = \frac{1}{\sqrt{2\pi}\sigma} \exp\left(-\frac{(R_i-\mu)^2}{2\sigma^2}\right),$$

$$p_{r_i|H_0}(R_i|H_0) = p_{n_i}(R_i) = \frac{1}{\sqrt{2\pi}\sigma} \exp\left(-\frac{{R_i}^2}{2\sigma^2}\right).$$

Поскольку n_i статистически независимы, совместная плотность вероятностей r_i является простым произведением индивидуальных плотностей вероятности. Таким образом, приведенные формулы можно записать в следующем виде:

$$p_{r|H_1}(R|H_1) = \prod_{i=1}^{N} \frac{1}{\sqrt{2\pi}\sigma} \exp\left(-\frac{(R_i-\mu)^2}{2\sigma^2}\right),$$

$$p_{r|H_0}(R|H_0) = \prod_{i=1}^{N} \frac{1}{\sqrt{2\pi}\sigma} \exp\left(-\frac{{R_i}^2}{2\sigma^2}\right).$$

Подставив полученные результаты в формулу коэффициента правдоподобия, получим

$$\Lambda(R) = \frac{\prod_{i=1}^{N} \frac{1}{\sqrt{2\pi}\sigma} \exp\left(-\frac{(R_i-\mu)^2}{2\sigma^2}\right)}{\prod_{i=1}^{N} \frac{1}{\sqrt{2\pi}\sigma} \exp\left(-\frac{{R_i}^2}{2\sigma^2}\right)}.$$

Сравнив полученное значение коэффициента правдоподобия с эмпирической шкалой, сделаем вывод о качестве модели в рамках виртуального эксперимента.

7 Описание архитектуры

Для проведения вычислительного эксперимента в рамках поставленных задач был собран программно-аппаратный вычислительный комплекс: аппаратная часть построена на базе серверов и коммутатора фирмы Quanta [23]; программная часть – программный кластер на базе продукта HDP (Hortonworks Data Platform) [16].

7.1 Аппаратная часть

Аппаратная архитектура состоит из шести серверов, представленных в форм-факторе OpenRack и объединенных одним коммутатором по интерфейсу 10Gb Ethernet. Каждый сервер имеет свои локальный SSD-диск с установленными на нём операционной системой и программными компонентами HDP кластера, а также свою полку с HDD-дисками, подключенную напрямую по интерфейсу SATA. Каждый сервер выполняет определенную роль в кластере, которая определяет его технические характеристики. Технические характеристики управляющих и рабочих узлов представлены в Таблице 4.

Таблица 4 Характеристики узлов

Характеристика	Значение
Тип процессора	Genuine Intel 2.30GHz/ Intel Xeon E5-2630L
Количество ядер	40/24
Память	32/64ГБ
Тип дисков	SSD, HDD
Объем дисков	240ГБ, 2/4ТБ
Резервирование дисков	RAID-1/JBOD
Подключение дисков	SATA
Операционная система	CentOS 6.9
Сетевые интерфейсы	10GbEthernet

Роли серверов в кластере:

- m1, m2 – управляющие узлы, к ним предъявляются повышенные требования в плане производительности и надежности; на них установлены компоненты, отвечающие за распределение задач по рабочим узлам кластера, такие, как Name Node сервиса HDFS, ресурс менеджер YARN и др.; критические компоненты зарезервированы в режиме Active-Passive; в качестве Active сервера для большинства из них выступает узел m1;
- s1, s2, s3, s4 – рабочие узлы; основной вычислительный элемент кластера обеспечивает выполнение программного кода приложений в распределенной среде; выход из строя такого узла не является критическим, так как их состояние постоянно отслеживается управляющими узлами, которые в случае падения перезапустят задачу на оставшихся доступных серверах, однако это приведет к замедлению выполнения расчетов на кластере.

Доступ к компонентам управления платформы осуществляется из общей сети лаборатории за счет

подключения коммутатора кластера с общим маршрутизатором по интерфейсу 1GbEthernet.

7.2 Программная часть

Для построение вычислительной системы была использована платформа HDP версии 2.6. Эта версия является последней актуальной на момент написания этой статьи. На аппаратные сервера был установлен минимально необходимый набор программных компонент для проведения виртуального эксперимента:

- HDFS [7] – отказоустойчивая распределенная файловая система;
- YARN [22] – менеджер ресурсов, основной планировщик задач, запускаемых на кластере;
- Storm [5], Spark (с модулем Streaming) [4], Flink [2] – системы потоковой обработки данных;
- Kafka [3] – виртуальная очередь;
- Ambari [16] – веб-интерфейс для управления кластером;
- ZooKeeper [17] – сервер координации работы распределенных приложений;
- Zeppelin [16] – среда написания и отладки программного кода.

Все представленные ниже компоненты, кроме Apache Flink, входят в установочный пакет HDP кластера. Flink установлен дополнительно как сервис над YARN. Установка одновременно трёх потоковых систем в рамках одного кластера обусловлена тем, что в настоящий момент нет единой методологии оценки. Поэтому типовым является сценарий, когда существующая практическая задача в тестовом виде реализуется одновременно на всех платформах, а уже затем сравнивается производительность, полученная на практике. Также на момент развертывания кластера были неизвестны ограничения всех систем, и соответственно не было принято решение о применимости конкретного продукта для реализации виртуального эксперимента.

8 Поток работы

Поток работ виртуального эксперимента представлен на Рис. 5.

Виртуальный эксперимент заключается в одновременной непрерывной работе следующих компонент над потоком входных данными:

- Producer – программа, занимающаяся извлечением данных из csv-файла и отправляющая сообщения в очередь Kafka;
- Kafka – обслуживает прием, промежуточное хранение, репликацию данных;
- Spark Streaming – выполняет роль Consumer, извлекает данные из очереди в формате RDD и передает их на обработку в ядро Spark.
- Spark – выполняет вычисления с данными, вызывает модель течения жидкости, проводит статистические тесты оценки качества модели.

Этапы потока работ таковы:

1. Данные, полученные в результате физического эксперимента, хранятся в распределенной файловой системе. Producer представляет собой программный модуль, реализованный на Python. Этот компонент первым шагом извлекает данные из CSV-файлов, содержащих показатели гидрофонов, значения температуры и скорости потока и формирует два массива строк;
2. Подготовка данных к отправке заключается в разбиении полученных массивов строк на партии по 100 тыс. значений для показателей гидрофонов и одного – для температуры и скорости потока;
3. Отправка сообщений осуществляется с помощью вызова метода send() из загруженной библиотеки kafka-python [19]. Этот метод является ассиметричным. Для увеличения скорости отправки значение Ack выставлено в 0, чтобы Producer не ждал от Kafka-пакета подтверждения доставки. Так как кластер находится в изолированном сетевом сегменте и на серверах используются высокоскоростные сетевые интерфейсы, то потерь пакетов не происходит;
4. Из-за ограничений Spark Streaming как потоковой системы [6], а именно:
    a) имеется только временное управление окном;
    b) реализованы чтение данных из всех разделов очереди и запись их в одну RDD;
    c) недостаточна производительность работы ядра Spark из-за использования промежуточного слоя Python-интерпретатора и самой архитектуры Spark;
    d) не реализован функционал Backpressure стандартными средствами;

    возникла необходимость в оптимизации гиперпараметров модели. Изначально предполагалось подавать на вход модуля обработки данных из очереди 100 тыс. сообщений в секунду. За данный интервал времени Spark должен их успевать обрабатывать. Однако, таких скоростей: как обработки, так и подачи в очередь достигнуть не удалось. Поэтому возникла необходимость в увеличении временного интервала отправки партии сообщений;
5. Слияние полученных признаков в один кортеж для отправки его на вход модели;
6. Расчет модели на базе полученных входных данных. Распределение задания расчета модели по кластеру выполняется ядром Spark, исходя из своей внутренней логики работы;
7. Проведение статистического теста. Для оценки качества модели мы используем одно-выборочный Т-тест с двух сторонней альтернативой или Байесовский подход. На

выходе мы имеем строку с результатом в формате:

a. число – для оценки в случае байесовского критерия;
b. True/False – для оценки статистического критерия;

8. Вывод результата в консоль.

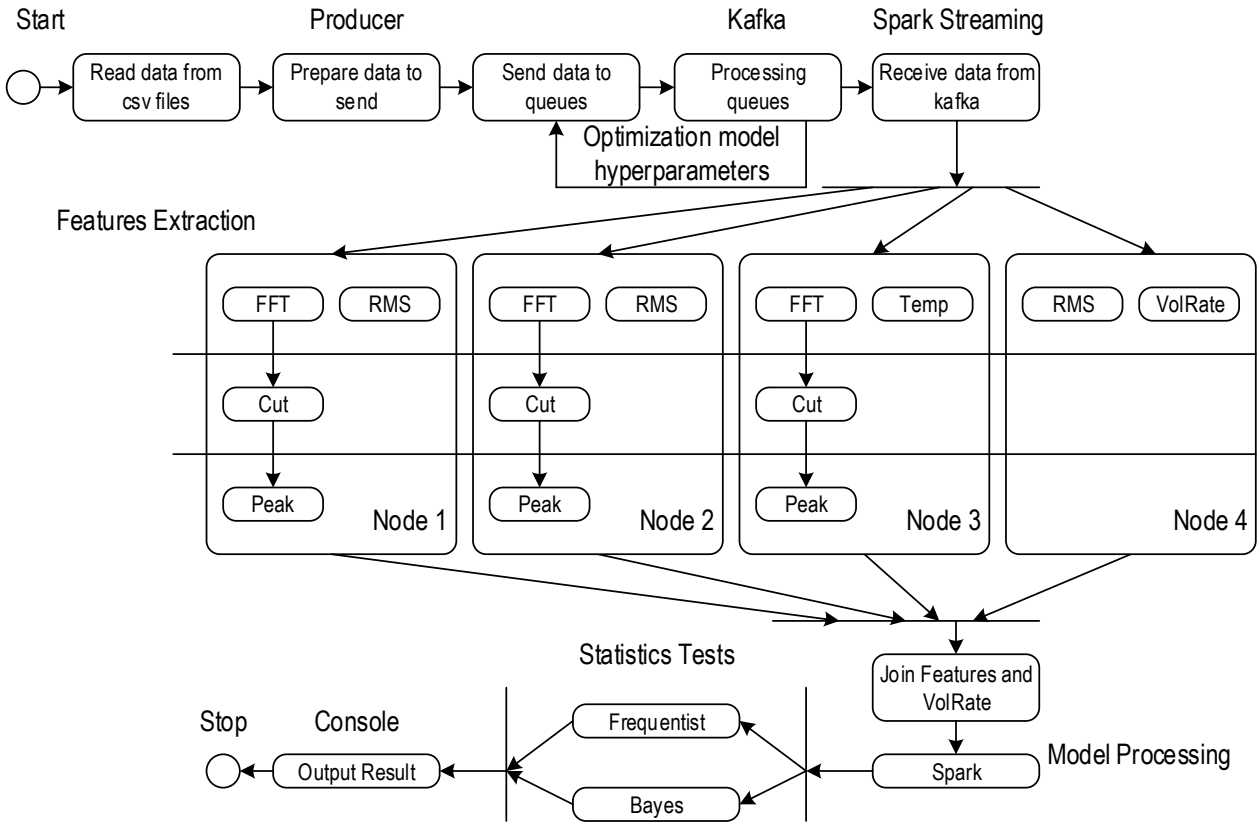


Рисунок 5 Поток работ

9 Полученный результат

По результатам проведенного эксперимента получены данные, представленные в Таблице 6. Цветом подсвечены результаты оценки качества виртуального эксперимента различными методами. Использованы следующие обозначения:

- белый – модель корректно описывает входные данные (Гипотеза H_0);
- серый – модель не корректна (Гипотеза H_1).

Таблица 6 Полученный результат

Байесовский	**Частотный**
0,0002	False
0,0296	False
0,6713	False
0,0053	False
0,0030	False
0,0006	False
1,1213	False
3,3046	True
5,4079	True
9,1870	True
20,0469	True
24,3965	True

По полученным результатам можно определить момент, когда модель начинает некачественно описывать поведение течения жидкости по входным данным. Результаты обоих методов схожи, но всё же немного отличаются, так как мы попали в граничные области Баейсовского критерия (3,3046), которые имеют слабую доказательную силу, поэтому отвергнуть гипотезу H_0 о том, что модель корректна, нельзя.

10 Заключение

Представлен подход, позволяющий оценить качество научных гипотез на примере области течения жидкости. Его идея базируется на классической теории детектирования, в рамках которой в качестве правил принятия решения выступают критерий Стьюдента и критерий правдоподобия.

Разработаны онтология предметной области и концептуальная схема виртуального эксперимента для исследования характеристик течения жидкости в трубе на основе его акустического шума. Произведен анализ существующих инструментов организации распределенной обработки потоковых данных.

Выполнено исследование подходов к организации методов проверки гипотез и оценки

качества моделей, реализующих соответствующие гипотезы.

Создана архитектура распределенной системы для оценки качества модели и проверки гипотез. Разработан масштабируемый программный модуль для существующей кластерной инфраструктуры.

Результаты эксперимента показали, что различные подходы проверки научных гипотез по оценке качества модели позволяют получить схожие результаты, незначительно отличающиеся на граничных областях критерия качества. Таким образом, частотный и Байесовский методы могут в равной степени быть применены для оценки качества виртуального эксперимента в рассматриваемой предметной области – течения жидкости.

Поддержка

Работа выполнена при поддержке РФФИ (грант 16-07-01028).

Литература

[1] Alexandrov, A., Bergmann, R., Ewen, S., Freytag, J., Hueske, F., Heise, A., Kao, O., Leich, M., Leser, U., Markl, V., Naumann, F., Peters, M., Rheinländer, A., Sax, M., Schelter, S., Höger, M., Tzoumas, K., Warneke, D.: The Stratosphere Platform aor Big Data Analytics. The VLDB J., 23 (6), pp. 939-964 (2014)

[2] Apache Flink: Scalable Batch and Stream Data Processing. https://flink.apache.org/

[3] Apache Kafka is a Distributed Streaming Platform. https://kafka.apache.org/intro

[4] Apache Spark – Lightning-Fast Cluster Computing. https://spark.apache.org/

[5] Apache Storm. https://storm.apache.org/

[6] Armbrust, M., Das, T., Davidson, A., Ghodsi, A., Or, A., Rosen, J., Stoica, I., Wendell, P., Xin, R., Zaharia, M.: Scaling Spark in the Real World: Performance and Usability. Proc. of the VLDB Endowment, 8 (12) (2015)

[7] Borthakur, D.: HDFS Architecture Guide. Hadoop Apache Project (2008)

[8] Brennen, C.E.: Fundamentals of Multiphase Flow. Cambridge University Press (2005)

[9] Chintapalli, S., Dagit, D., Evans, B., Farivar, R., Graves, T., Holderbaugh, M., Liu, Z., Nusbaum, K., Patil, K., Peng, B., Poulosky, P.: Benchmarking Streaming Computation Engines: Storm, Flink and Spark Streaming. Proc. of the IEEE Int. Parallel and Distributed Processing Symposium Workshops (2016)

[10] Duggan, J., Brodie, M.: Hephaestus: Data Reuse for Accelerating Scientific Discovery. Proc. of 7th Biennial Conf. on Innovative Data Systems Research (CIDR'15). USA (2015)

[11] Goncalves, B., Porto, F.: Managing Large-Scale Scientific Hypotheses as Uncertain and Probabilistic Data With Support for Predictive Analytics. Proc. of the IEEE Computing in Science and Engineering, 17 (5), pp. 35-43 (2015)

[12] Goncalves, B., Silva, F., Porto, F.: ϒ-DB: A System for Data-Driven Hypothesis Management and Analytics (2014). http://arxiv.org/abs/1411.7419

[13] Hagedorn, S., Götze, P., Saleh, O., Sattler, K.: Stream Processing Platforms for Analyzing Big Dynamic Data. Information Technology, 58 (4), pp. 195-205 (2016)

[14] Harry, L. van Trees, Bell, K., Tian, Z.: Detection, Estimation, and Modulation Theory. Part 1 - Detection, Estimation, and Filtering Theory. Second Edition. Willey, 1175 p. (2013)

[15] Hesse, G., Lorenz, M.: Conceptual Survey on Data Stream Processing Systems. Proc. of the IEEE 21st Int. Conf. on Parallel and Distributed Systems (2015)

[16] Hortonworks Data Platform. https://hortonworks.com/products/data-center/hdp/

[17] Hunt, P., Konar, M., Junqueira, F., Reed, B.: ZooKeeper: Wait-free Coordination for Internet-Scale Systems. Proc. of the USENIX Annual Technical Conf. (2010)

[18] Ioannidis, J.P.: Why Most Published Research Findings Are False. PLoS Medicine, 2 (8) (2005)

[19] Kafka-python. http://kafka-python.readthedocs.io/en/master/index.html

[20] Kalinichenko, L., Kovalev, D., Kovaleva, D., Malkov, O. Methods and Tools for Hypothesis-driven Research Support: a Survey. Informatica and Appications, 9 (1), pp. 28-54 (2015)

[21] Marz, N., Warren, J.: Big Data: Principles and Best Practices of Scalable Realtime Data Systems. Manning Publications Co. USA. 1st edition (2015)

[22] Mathiya, B., Desai, V.: Apache Hadoop Yarn Parameter Configuration Challenges and Optimization. Proc. of the Int. Conf. on Soft-Computing and Network Security (ICSNS -2015). Coimbatore. India (2015)

[23] Quanta Cloud Technology. http://qct.io/

[24] Rysak, A., Litak, G., Mosdorf, R.: Analysis of Non-stationary Signals by Recurrence Dissimilarity. Recurrence Plots and Their Quantifications: Expanding Horizons (2016)

[25] Salvetti, M.V., Geurts, B., Meyers, J., Sagaut, P.: Quality and Reliability of Large-Eddy Simulations. Springer. Netherlands (2008)

[26] Schales, D., Hu, X., Jang, J., Sailer, R., Stoecklin M., Wang, T.: FCCE: Highly Scalable Distributed Feature Collection and Correlation Engine for Low Latency Big Data Analytics. Proc. of 2015 IEEE 31st Int. Conf. on Data Engineering, pp. 1316-1327 (2014)

[27] Spark documentation. Pyspark.streaming module. https://spark.apache.org/docs/1.6.3/api/python/pyspark.streaming.html

[28] Spark Programming Guide. https://spark.apache.org/docs/1.6.3/programming-guide.html

[29] Stonebraker, M., Çetintemel, U., Zdonik, S.: The 8 Requirements of Real-time Stream Processing. SIGMOD Rec., 34 (4), pp. 42-47 (2005)

[30] Ünalmis, Ö.H.: Subsea Multiphase Flowmeter: Performance Tests in Multiphase Flow Loop. Society of Petroleum Engineers

[31] Wilcox, D.C.: Turbulence modeling for CFD. D C W Industries (2006)

[32] Zagarola, M.V., Smits, A.J.: Mean-flow Scaling of Turbulent Pipe Flow. J. of Fluid Mechanics, 373, pp. 33-79 (1998)

[33] Байесовский подход к теории вероятностей. Примеры Байесовских рассуждений. Глава 6. 2007. http://www.machinelearning.ru/wiki/images/4/43/BayesML-2007-textbook-2.pdf

[34] Критерий Стьюдента. http://www.machinelearning.ru/wiki/index.php?title=Критерий_Стьюдента

[35] Проверка статистических гипотез. http://www.machinelearning.ru/wiki/index.php?title=Проверка_статистических_гипотез

[36] Самарев, Р.С.: Обзор состояния области потоковой обработки данных. Труды ИСП РАН, (1), сс. 231-260 (2017)

[37] Скворцов Н.А., Аввакумова, Е.А., Брюхов, Д.О., Вовченко, А.Е., Вольнова, А.А., Длужневская, О.Б., Кайгородов, П.В., Калиниченко, Л.А., Князев, А.Ю., Ковалева, Д.А., Малков, О.Ю., Позаненко, А.С., Ступников, С.А.: Концептуальный подход к решению задач в астрономии. Астрофизический бюллетень, 71 (1), сс. 122-133 (2016)

[38] Тарасов, Е.А.: Сокращение числа виртуальных экспериментов с помощью оценки корреляций параметров взаимодействующих гипотез. Сб. трудов XVIII Межд. конф. DAMDID/RCDL, сс. 383-388 (2016)

Organization of Virtual Experiments in Data-Intensive Domains: Hypotheses and Workflow Specification

© Dmitry Kovalev © Leonid Kalinichenko © Sergey Stupnikov

Federal Research Center «Computer Science and Control» of Russian Academy of Sciences, Moscow, Russia

dkovalev@ipiran.ru lkalinichenko@ipiran.ru sstupnikov@ipiran.ru

Abstract. Organization and management of virtual experiments in data-intensive research has been widely studied in the several past years. Authors survey existing approaches to deal with virtual experiments and hypotheses, and analyze virtual experiment management in a real astronomy use-case. Requirements for a system to organize virtual experiments in data intensive domain have been gathered and overall structure and functionality for system running virtual experiments are presented. The relationships between hypotheses and models in virtual experiment are discussed. Authors also illustrate how to conceptually model virtual experiments and respective hypotheses and models in provided astronomy use-case. Potential benefits and drawbacks of such approach are discussed, including maintenance of experiment consistency and shrinkage of experiment space. Overall, infrastructure for managing virtual experiments is presented.

Keywords: virtual experiment, hypothesis, conceptual modeling, data intensive domains.

1 Introduction

Data intensive research (DIR) is evolving according to the 4th paradigm of scientific development and reflects the fact that modern science is highly dependent on knowledge extraction from massive datasets [5]. Data intensive research is multidisciplinary in its nature, bringing in many separate principles and techniques to handle complex data analysis and management. Up to 80% of researcher's time is spent on management of raw and analytical data, including data collection, curation and integration. The rest part requires knowledge inference from collected data in order to test proposed hypotheses, gather novel information and correctly integrate it. Although, it is the core of scientific work, it takes just 20% of researcher's time. To overcome that, a new approach for handling multidisciplinary DIR is needed.

Large-scale scientific experiments besides data processing issues are highly sophisticated– they include workflows, models and analytical methods. Every implementation of DIR can be treated as virtual experiment over massive collections of data. In [7] a survey is presented discussing different approaches to experiment modeling and how its core artifacts – hypotheses, can be specified. The use of conceptual representation of hypotheses and their corresponding implementation is emphasized, thus leading to the need of proper tools.

The article aims at developing methods and tools to support the execution and conceptual modeling of virtual experiment and designing infrastructure to manage it.

Article is structured as follows. In Section 2 related works are discussed. Section 3 explains why systems from section 2 are not enough, and introduces real-world use-case coming from astronomy. In section 4 main notions are defined. Section 5 provides infrastructure and functionality of system components is proposed. Section 6 concludes the article.

2 Related works

Systems with explicit representation of hypotheses are being rapidly developed during last several years [2–4, 6, 10]. Authors analyzed 3 different systems for executing virtual experiments and hypotheses: Hephaestus, Upsilon-DB and SDI. Some requirements for organizing and managing virtual experiments were extracted during the analysis. Although these platforms provide some important insights into defining and handling hypotheses, they miss some important features. First, they do not describe the perception of automatically derived hypotheses by domain experts, do not track their evolution, and do not discuss experiment design principles.

Hephaestus. It is a system for running virtual experiments over existing collections of data. It provides independence from resources and the system rewrites its queries into data source queries. System hides underlying implementation details from user, letting him work only with Hephaestus language. The language itself is a SQL-like language and is used to specify virtual experiment and underlying hypotheses.

Hephaestus separates two different classes of hypotheses: top-down and bottom-up. Top-down hypotheses are the one introduced by the researcher, while bottom-up hypotheses are derived from data. System supports the discovery of bottom-up hypotheses by looking for the correlation in data. These hypotheses are then ranked by some score (e. g. p-value of some statistical test) and the one with highest are passed to the researcher. Yet the system does not support automatical finding of causality, which is an important requirement for the future work. Hephaestus emphasizes the role of

Proceedings of the XIX International Conference "Data Analytics and Management in Data Intensive Domains" (DAMDID/RCDL'2017), Moscow, Russia, October 10–13, 2017

the expert in understanding which relationships should be further studied and which should not be chased. Hephaestus also computes metrics about experiments to estimate significance adequate to abandon further computation. System is used in testing clinical trials. The system does not catch the evolution of hypotheses or experiments yet.

Upsilon-DB. System enables researcher to code and manage deterministic scientific hypotheses as uncertain data. It uses internal database to form hypotheses as relations and adds uncertainty parameter. Later, that uncertainty parameter is used to rank hypotheses using Bayes rule. Provided approach can be treated as complementary to classical statistical approach. The systems allows to work with two types of uncertainty - theoretical, which is brought by competing hypotheses, and empirical uncertainty, which appears because of alternative datasets used. The system introduces algorithm to rank hypotheses using observed data. This is done because several competing hypothesis can explain the same observation well and some score to distinguish them is needed. When new data becomes available, this score can be adjusted accordingly.

Hypotheses have mathematical representation and authors provide method to translate its mathematical representation into relations in database. The simulations are also treated as data and respective relations are put inside the same database as hypotheses. Authors emphasize the need to support and develop the extraction of hypotheses from data and methods to sample both hypotheses and data. They illustrate that systems such as *Eureqa* [8] can be used to learn formula representation from data.

Following example is presented in the paper: authors present three different laws describing free fall and some simulated data. They rank hypotheses accordingly.

SDI. Platform is used to support scientific experiments. The system has the ability to integrate open data, reuse observed data and simulation data in the further development of experiments. The system enables multiple groups of researchers to access data and experiments simultaneously. Components of the framework are developed in such way that they could be deployed, adapted and accessed in individual research projects fast. SDI requires the support of lineage, provenance, classification, indexing of experiments and data, the whole cycle of obtaining data, curating and cleaning it, building experiments to test hypotheses over massive data, aggregating results is supported over long periods of time. The use of semantics is required by the system.

3 Astronomical Use-case

Surveyed systems do not cover several important issues, including interaction between hypotheses in single experiment, tracing experiment evolution, perception of automatically derived hypotheses and formulas by field experts.

Authors' further experience on how to deal with virtual experiments and hypotheses is based on *Besancon Galaxy Model (BGM)*. BGM is based on "the population synthesis approach … aims at assembling current scenarios of galaxy formation and evolution, theories of stellar formation and evolution, models of stellar atmospheres and dynamical constraints, in order to make a consistent picture explaining currently available observations of different types (photometry, astrometry, spectroscopy) at different wavelengths".

BGM which is being developed for more than 35 years represents a complex computational artifact, described in a series of [1, 11, 12]and presented in several major releases. Such a development represents a unique experience for catching the evolution scenarios for the model, changes to the model introduced both by using new observations (e.g. Hypparcos and Tycho-2 surveys) and the theoretical progress in the field. Both small changes to parameters of the model and huge improvements of the whole process were also made during the lifetime of the model. Also, the BGM authors enabled the community to change some parts of the model.

Due to the great experience collected by the BGM authors in the respective articles and associated code, now there is a possibility to collect the requirements for the system to supports experiments and provide rationale to choosing the appropriate methods and adequate techniques for the infrastructure.

BGM takes as input hypotheses and their parameters. The examples of such hypotheses are star formation rate (SFR), initial mass function (IMF), density laws, evolutionary tracks and so on [1]. As the model is evolving, new values for hypotheses parameters, even new parameters have been introduced into the BGM, e.g. for the IMF hypothesis in the last realization there has not only been tests of several new values of the hypothesis, but also separation of 2-slope and 3-slope instances of IMF is done.

It is very important to explicitly catch the relationship between several hypotheses in VE. Hypotheses and their parameters can be interrelated. For example, stellar birthrate function is derived from both IMF and SFR functions and local volume density function is based on provided density law. The relationships between hypotheses put constraints on the tuning of their parameters – model can quickly become.

Parameters of a single hypothesis can be linked to each other directly through equations. There are also indirect connections of parameters of several hypotheses, e.g. SFR parameter correlates with the slopes of IMF. This implies that one could not give the best solution for a particular variable without correlating it with others. So, there is a need to support for a correlation search between hypotheses parameters and to store relationships between parameters of a single hypothesis.

Not all model ingredients are allowed to be changed by the user. This is done because if some hypothesis is changed in the model and no further adjustments for the dependant hypotheses are made a model consistency is

broken. Furthermore, the model has a property of being self-consistent meaning that when input values change, if it is possible hypotheses derived by the one changed are properly adjusted in order not to break fundamental equations of astronomy. Therefore, derived by relationship needs to be modeled. Also, system component should enable the adjustment and calibration of any hypothesis available in the model.

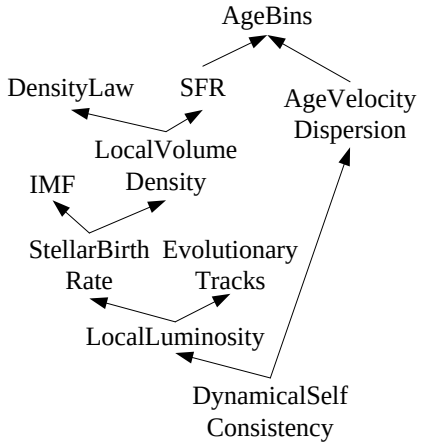


Figure 1 BGM Hypotheses Lattice with derived by relationship

Apart from explicit hypotheses, there are also implicit hypotheses in the model. They are not described in the articles and are tacit. The example of such hypothesis is that no stars come from outside of the Galaxy. It is important to explicitly store such hypotheses and understand how to extract such hypotheses from publications and data sources.

Workflow is used to implement BGM experiment specifying when each model which conforms to related hypotheses should be invoked. The workflow has also evolved since the first version, e.g. for thin disk treatment new activities dependent on IMF and SFR hypotheses are introduced. This development can only be tracked using publications. Some activities in model structure require the usage of statistical methods, tests and tools, which are used on both local hypotheses and on the general simulations from the whole experiment.

As the number of experiments is huge due to the increasing size of competing hypotheses family, now not all of the possible are run against the whole sky. Studying the ways to reduce the number of experiments which give the best fit and to choose when and if to abandon further computations of experiment is a major part of requirements to the new system. Using the information from experiment run done both locally and by other research groups can be helpful in achieving that goal.

Some researches of data-intensive analysis emphasize the role of error bars. As the data in astronomy is provided usually with errors, the BGM uses special methods to work with such type of uncertainty. A component supporting statistical tools which works with error bars is a major requirement for the infrastructure.

4 Hypotheses and Models in Virtual Experiment

4.1 Main Notions

Extracted information needs to be formally specified. For that, authors define additional artifact – virtual experiment. It is a tuple **<O, H, M, R, W, C>**, where **O** is a domain ontology. Domain ontology is a set of concepts and relationships in applied domain formally specified with some language.

H is a set of hypotheses specifications and relationships between them. H is a part of ontology and uses concepts from it. Together they form the ontology of virtual experiment. Hypothesis is a proposed explanation of a phenomenon that still has to be rigorously tested.

M is a set of models. Each model is a set of functions. Every model implements a hypothesis specification. If model generates expected behavior of some phenomenon, it is said that model and respective hypothesis are supported by observations.

R: H -> M, is a mapping from the set of hypotheses and into the models.

W is a workflow. Workflow is a set of tasks, orchestrated by specific constructs (workflow patterns - split, join, etc.). Each task represents a function with predefined signature, which invokes models from M. Workflow implements experiment specifying when each model that conforms to related hypotheses should be invoked.

C is a configuration for each experiment run. It consists of a total mapping from workflow tasks into sets of function parameter values.

There exist a lot of possible hypotheses representations – mathematical models, Boolean networks, ontologies, predicates in first-order logic, etc. Authors use ontologies to specify hypotheses.

Possible relationships between hypotheses are *competes_with*, which is used to relate competing hypotheses and *derived_by* to relate two hypotheses, one of which was used to derive another. *Derived_by*can be used to form hypotheses lattice [9] – algebraic structure with partial order relation. Hypotheses derived from a single hypothesis are atomic, otherwise – complex (see Fig. 1).

Model, which implements hypothesis, should conform to the hypothesis specification. If model generates expected behavior of some phenomenon, it is said that model and respective hypothesis are supported by observations.

4.2 Remarks on methodology

Since hypotheses become the core artifact of virtual experiment, there is a shift in treating data to successfully manage it. Fig. 2 depicts the process of specifying virtual experiment.

First, hypotheses are extracted from articles. Usually, it is text or formulas. Sometimes, there is a need to provide external hypotheses and substitute existing ones. Next step is to define mapping between hypotheses and models, which implement these hypotheses, and build some workflow specifying the sequence of tasks. Forming a research lattice is a next step. Virtual experiment needs configuration and execution plan. After that, one can launch virtual experiment.

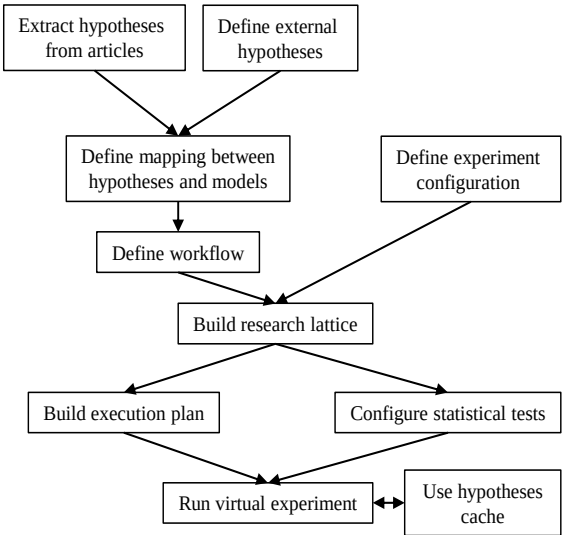


Figure 2 Methodology to form virtual experiment

4.3 Virtual Experiment Specification

Conceptual schema to define virtual experimentis provided. It is written with the simplified OWL functional syntax (*Declaration* keyword is omitted; property, domain, and range declarations are combined).Virtual experiment (*VirtualExperiment* class) has associated set of *hypotheses*, single *workflow*, *observed_data* against which experiment will run and *probability*, which describes how well underlying model suits observed data. Closer probability is to 1, better the underlying model simulates phenomenon.

```
Ontology(<http://synthesis.ipi.ac.ru/virtual_experiment/ontology>

Class(VirtualExperiment)

ObjectProperty(Hypothesis

domain(VirtualExperiment) range(Hypothesis))

ObjectMinCardinality(1 Hypothesis
     VirtualExperiment)

DataProperty(workflow domain(VirtualExperiment)
range(xsd:anyURI))

DataMinCardinality(1 workflow
     VirtualExperiment)

DataProperty(mediatordomain(VirtualExperiment)
range(xsd:anyURI))

DataMinCardinality(1 mediator
     VirtualExperiment)

     DataProperty(probability

       domain(VirtualExperiment)
range(xsd:float))

     DataExactCardinality(1          probability
     VirtualExperiment))
```

Hypothesis is specified in the same ontology as virtual experiment. Every hypothesis has *name*, *description*, *author(s)* and associated *articles*. It also has a model associated with it. Following [4] associated *probability* of hypothesis is introduced.

Several hypotheses explaining one and the same phenomena are called *competing*. Also hypothesis can be derived by some other hypothesis. Hypotheses lattice is formed with *derived_by* relationship on hypotheses space.

```
Class(Hypothesis)

DataProperty(probability domain(Hypothesis)
     range(xsd:float))

DataExactCardinality(1probabilityHypothesis)

DataProperty(name domain(Hypothesis)
     range(xsd:string))

DataExactCardinality(1name   Hypothesis)

DataProperty(description domain(Hypothesis)
     range(xsd:string))

DataMinCardinality(1descriptionHypothesis)

DataProperty(author
domain(Hypothesis)range(xsd:string))

DataMinCardinality(1authorHypothesis)

DataProperty(article domain(Hypothesis)
     range(xsd:anyURI))

DataMinCardinality(1article    Hypothesis)

DataProperty(model domain(Hypothesis)
Hypothesis)

DataExactCardinality(1 model range(xsd:anyURI))



Class(HypothesisMetaClass)

ClassAssertion(HypothesisMetaClassHypothesis)

ObjectProperty(competes
domain(HypothesisMetaClass)

range(HypothesisMetaClass))

ObjectProperty(derivedBydomain(HypothesisMetaClass)

range(HypothesisMetaClass))
```

4.4 Hypotheses Specification

Examples of hypotheses and their relationships come from Besancon Galaxy Model (BGM). For the sake of clarity not all hypotheses in BGM are specified. All of the BGM hypotheses are treated as subclasses of Hypothesis class.

Initial Mass Function is the mass distribution of a given population of stars and is represented by standard power law. Due to construction of the hypothesis in the BGMIMF has a mathematical representation as a piecewise function with 2 or 3 pieces (slopes) where it is defined for mass regions. As there are just 2 possible sizes of the piecewise function, we put this into two disjoint subclasses. There are restrictions on available mass to Sol mass ratio. For IMF, authors test 10 different

versions of a hypothesis, 4 of them are 2-slope functions and 6 of them are 3-slope function. All of tested hypotheses are competing. Example instance from each subclass is given.

```
Class(Slope)

DataProperty(alpha domain(Slope)
range(xsd:float))

DataProperty(minMass domain(Slope)
range(xsd:float))

DataProperty(maxMass domain(Slope)
range(xsd:float))

DataExactCardinality(1 alpha Slope)

DataExactCardinality(1 minMassSlope)

DataExactCardinality(1 maxMassSlope)

SubClassOf(IMF Hypothesis)

ObjectProperty(Slopes domain(IMF) range(Slope))

DataProperty(availableMass domain(IMF)
     range(xsd:float))

DataExactCardinality(1 availableMass  IMF )

DataProperty(outputStarMass domain(IMF)
     range(xsd:float))

DataExactCardinality(1 outputStarMass IMF )

SubClassOf(ThreeSlopeIMF IMF)

ObjectExactCardinality(3 Slopes  ThreeSlopeIMF)

SubClassOf (TwoSlopeIMF IMF)

ObjectExactCardinality(2 Slopes  TwoSlopeIMF )

DisjointClasses(TwoSlopeIMFThreeSlopeIMF)


ObjectPropertyAssertion(competes TwoSlopeIMF
IMF)

ObjectPropertyAssertion(competes ThreeSlopeIMF
IMF)

ClassAssertion(Slope HaywoodSlope1)

DataPropertyAssertion(alpha HaywoodSlope1
"1.7"^^xsd:float)

DataPropertyAssertion(minMass HaywoodSlope1
"0.09"^^xsd:float)

DataPropertyAssertion(maxMass HaywoodSlope1
"1.0"^^xsd:float)

ClassAssertion(Slope HaywoodSlope2)

DataPropertyAssertion(alpha HaywoodSlope2
"2.5"^^xsd:float)

DataPropertyAssertion(minMass HaywoodSlope2
"1.0"^^xsd:float)

DataPropertyAssertion(maxMass HaywoodSlope2
"3.0"^^xsd:float)

ClassAssertion(Slope HaywoodSlope3)

DataPropertyAssertion(alpha HaywoodSlope3
"3.0"^^xsd:float)

DataPropertyAssertion(minMass HaywoodSlope3
"3.0"^^xsd:float)

DataPropertyAssertion(maxMass HaywoodSlope3
"120.0"^^xsd:float)

ClassAssertion(ThreeSlopeIMFHaywoodIMF)

ObjectPropertyAssertion(Slopes
HaywoodIMFHaywoodSlope1)

ObjectPropertyAssertion(Slopes HaywoodIMF
HaywoodSlope2)

ObjectPropertyAssertion(Slopes HaywoodIMF
HaywoodSlope3)
```

Star Formation Rate, $\Psi(t)$ represents the total mass of stars born per unit time per unit mass of Galaxy. Star formation rate has subclasses for representing constant $\Psi(t) = C$ and exponential function $\Psi(t) = \exp\{-\gamma t\}$ where γ is a parameter. Authors tested several competing hypotheses - two possible values for gamma (0.12 and 0.25) and one constant value. They can be stated as instances of respective classes.

```
SubClassOf(SFR Hypothesis)

DataProperty(time domain(SFR) range(xsd:float))

DataExactCardinality(7 time  SFR )

DataProperty(bornStarMass domain(SFR)
     range(xsd:float))

DataExactCardinality(7 bornStarMass SFR )

SubClassOf(ConstantSFR SFR)

DataProperty(constant domain(ConstantSFR)
range(xsd:float))

DataExactCardinality(1 constant )

SubClassOf(ExponentSFR SFR)

DataProperty(gamma domain(ExponentSFR)
range(xsd:float))

DataExactCardinality(1 gamma)

DisjointClasses(ExponentSFRConstantSFR)

ObjectPropertyAssertion(competes ConstantSFR
SFR)

ObjectPropertyAssertion(competes ExponentSFR
SFR)

ClassAssertion(ExponentSFRRobinSFR)

DataPropertyAssertion(gamma RobinSFR
     "0.12"^^xsd:float)
```

BGM apart from model ingredients has also implicit hypotheses, which are not marked as ingredients. For example, 1) thin disk is divided into seven age bins; 2) no stellar population comes from the outside of the galaxy. For the first example we can specify additional class AgeBins which has exactly seven age bins.

```
SubClassOf(AgeBins Hypothesis)

DataProperty(ageBin domain(AgeBins)
range(xsd:integer))

DataExactCardinality(7 ageBin  AgeBins)
```

It is more difficult to deal with the second one. As a possible solution, additional hypothesis could later be specified.

Hypotheses lattice is modeled with derived Byobject property. Some classes can be specified using Equivalent Classes construction. Hypotheses lattice for BGM was created manually, but later it should be constructed automatically by system for executing experiments. (Part of) hypotheses lattice for BGM is shown in Fig. 1.

```
ObjectPropertyAssertion(derivedBy SFR AgeBins)

ObjectPropertyAssertion(derivedByAgeVelocityDis
persionAgeBins)

ObjectPropertyAssertion(derivedBy SFR
LocalVolumeDensity)

ObjectPropertyAssertion(derivedByDensityLawLoca
lVolumeDensity)
```

For IMF class and there are relations between slopes, output Mass and available Mass. Based on available Mass parameter alpha is chosen and then output Mass is computed. If available Mass is inside the respective interval, alpha is taken and output Mass is computed. Next, post-condition for ExponentSFR is written. It says that born stars should have mass respective to the exponential equation. Other pre- and post-conditions are specified in the same manner.

```
Document(

Group(Forall ?IMF  ?am ?s ?om ?a ?min ?max (

AND (?IMF[AvailableMass -> ?am Slopes -> ?s
outputStarMass ->

    ?om]  ?s[alpha -> ?a minMass -> ?min
maxMass -> ?max]) :-

  AND (External(pred:numeric-greater-than(?am

External(func:numeric-
multiply(?mincon:solMass)))

External(pred:numeric-less-than(?am
External(func:numeric-

multiply(?max con:solMass))))

))))

Forall ?ExponentSFR ?g ?t ?m(

?ExponentSFR[Gamma -> ?g Time->?t BornStarMass-
> ?m]:- AND(

External(pred:numeric-equal(?m

External(func:numeric-exponent(func:numeric-
multiply(

"-1.0"^^xsd:float?t)?g)))))))
```

4.5 Workflow Specification

The model of mass determination consists of a local mass normalization, the simulation of the local neighborhood and calculating vertical density distribution. These tasks can be further divided into several subtasks:

1. *getRSVDensity*. Relative density is calculated using Einasto density law. After that for each population this density $\rho(r, l, b, i)$ is integrated in the vertical direction ratio of surface to volume density (RSV) is computed.
2. *getSurfaceDenisty*. For each thin disk subcomponent surface density is calculated and then summed. Surface density of each age subcomponent has to be proportional to the intensity of SFR in its respective age bin.
3. *getVolumeDensity*. Volume stellar mass densities are calculated and summed Total volume is checked to fit the observations.
4. *adjustSurfaceDensity*. If the difference occurs surface and volume density are adjusted and recomputed.
5. *getLNSimulations*. Provided with specific hypotheses (IMF, SFR, evolutionary tracks and so on) stars and their parameters are simulated in the local neighborhood.
6. *getAliveStarsRemnants*. Stars are splitted into alive stars and remnants. Remnants -are possible stars for which the age and mass combination was not on the evolutionary tracks.
7. *solvePotentialEquation*. Poisson equation is solved with the input of stellar content of thin disk.
8. *constrainPotential*. Calculated potential should be constrained by observed Galactic rotation curve. The central mass and corona parameters are computed in such a way that the potential reproduces the observed rotation curve.
9. *calculatePotentialParameters*. Based on the calculated potential central mass parameters and corona parameters are computed.
10. *solveBoltzmannEquation*. Boltzmann collisionless equation for an isothermal and relaxed stellar population is solved in order not to break fundamentals of the model.
11. *checkDynamicalConsistency*. As equations in 6,7,8 are solved separately, the potential does not satisfy both constraints. These tasks should be run until the changes in the potential and other parameters are less than 0.01.

Workflow is specified as a RIF-PRD document. The ontology for virtual experiment and BGM ontology are imported. Rules in the document are separated into two groups. The first group with priority 2 is used to define workflow input and output parameters and variables. Part of specification describes several hypotheses passed as input parameters and calculated local surface density for each age bin as output. GetLocalSurfaceDensity task is specified in a group with priority 1. Task gets as input SFR hypothesis and total surface density vector (initially a guess) and multiplies provided values. Task checks if Xor of dependent tasks is done.

```
Document( Dialect(RIF-PRD)

Base(<http://synthesis.ipi.ac.ru/virtualexperim
ent/workflow#>)

Import(<http://synthesis.ipi.ac.ru/virtualexper
iment/ontology#>)

Import(<http://synthesis.ipi.ac.ru/bgm/ontology
#>)Prefix(bgm<http://sy

nthesis.ipi.ac.ru/bgm/ontology#>)

Prefix(ve<http://synthesis.ipi.ac.ru/virtualexp
eriment/ontology#>)

Group 2 (

Do(

Assert(External(wkfl:parameter-
definition(sfrbgm:SFRIN)))

Assert(External(wkfl:parameter-
definition(imfbgm:IMF IN)))

Assert(External(wkfl:parameter-definition(avd
```

```
bgm:AgeVelocityDispersionIN)))

Assert(External(wkfl:parameter-definition(dl

bgm:DensityLaw IN)))

Assert(External(wkfl:parameter-definition(et

bgm:EvolutionaryTracks IN)))

Assert(External(wkfl:variable-
definition(lsdsList(xsd:float)

IN)))

Assert(External(wkfl:variable-definition(clsds
List(xsd:float)

      OUT)))

Assert(External(wkfl:variable-value(clsd
List())))

Group 1 (

Do(

Forall ?sfr?bsm?lsd ?lsds ?clsd ?clsds such
that (

External(wkfl:variable-value(lsds ?lsds))

External(wkfl:variable-value(clsds ?clsds))

External(wkfl:variable-value(sfr ?sfr))

      ?lsd#?lsds

      ?clsd#?clsds

      ?sfr[bornStarMass -> ?bsm]

( IfOr(Not(External(wkfl:end-of-
task(getRSVDensity)))

External(wkfl:end-of-
task(adjustSurfaceDensity)) )

Then Do( Modify(?clsd ->External(func:numeric-
multiply(?bsm

      ?lsd)) )

Assert(External(wkfl:end-of-
task(getSurfaceDensity))) ))
```

4.6 Choosing parameters of hypotheses for virtual experiment execution

Since some hypotheses can take quite a few values, the number of possible models can reach thousands. This poses a question about the order of model execution and how to make these executions effective (and not to recompute previous unchanged results). For that we use special structures to cache and store results. The system can put model execution in some order and use the results of previous executions. This could drastically increase the speed of model computation, especially on big amount of data. To implement this we use properties of hypotheses lattices.

The researcher can run several experiments finding the probability of each, which can be later queried by other researchers. For example, following query takes two experiments, which have underlying models best explaining observed data, and fixed values for hypothesis SFR and workflow specified by URI. Since there could be thousands of possible experiments, there is a need to order them by their probability. As in [3] we don't want the researched to bury in thousands of possible models and just take several best ones.

```
SELECT ?experiment

WHERE {

     ?experiment probability ?probability .

     ?experiment workflow ?workflow .

     ?experiment Hypothesis ?hypothesis .

     ?hypothesis name ?name .

     FILTER(?name = 'RobinSFR' && ?workflow =
URI)

}

ORDER BY desc(?probability)

LIMIT 2
```

5 Requirements for Infrastructure for Managing Virtual Experiments

In a series of experiment run it is important to keep track on evolution of models, hypotheses and experiments, as well as identifying new data sources.

Operations to manipulate virtual experiments and its components need to be defined. Next, the system needs to capture dependencies (competes, derived by) between hypotheses, invariants in single hypothesis. Correlations between parameters of several hypotheses should also be considered.

Second, infrastructure should contain components responsible for automatic extraction of dependencies between hypotheses, parameters in single and multiple hypotheses. Obtained data is used in deciding which experiments should be abandoned and also used in keeping hypotheses in a single experiment consistent.

Third, one needs components for maintaining experiment consistency and constraining the number of possible experiments as well as defining the metric which is used to define if experiment poorly explains phenomena and abandon further computations. Methods for removing poor experiments based on previous experiments runs are also required. Experiments and hypotheses should stay consistent when parameters of a hypotheses change.

As soon as several hypotheses in some experiments could explain some phenomena well and due to errors in data, researcher needs to deal with uncertainty and needs methods to rank experiments and competing hypotheses on massive datasets.

While experiment could change slightly from a previous experiment run (e. g. one hypothesis parameter changes), system should store some data about previous executions. Methods for understanding which parts of experiments should be recomputed and which are not should be developed as well. Defining structures to store results of previous experiments and query these results is important. Since there could be thousands of possible experiments system should use a method to form a plan to execute experiments in such way that stored results are mostly used and no additional recomputations are made.

Some stages will investigate and adopt or reject certain values such as a velocity hypothesis, then continue. The design of the paths to be followed is called experimental design that as in the scientific method is the hardest part of the analysis. In principle, as in many systems, Hephaestus could pursue multiple paths in parallel using some metric to determine when to abandon a path. Some have criticized DeepDive and others for following a single path.

Reducing computational experiments (what we call virtual experiments), as mentioned above using metrics to estimate significance adequate to abandon further computation.

6 Conclusion

The article aims at developing a new approach to managing virtual experiment. Hypotheses are becoming core artifacts of that approach. By analyzing existing systems and use case requrements are extracted. Formal specification of the determination of the mass model from BGM is presented in the OWL syntax.

Further work should be concentrated on developing metasystem for handling hypotheses, models and other metadata in virtual experiment.

Acknowledgments

This research was partially supported by the Russian Foundation for Basic Research (projects 15-29-06045, 16-07-01028).

References

[1] Czekaj, M. et al.: The Besancon Galaxy Model Renewed I. Constraints on the Local Star Formation History from Tycho Data. *arXiv preprint arXiv:1402.3257*. (Feb. 2014)

[2] Demchenko, Y. et al.: Addressing Big Data Issues in Scientific Data Infrastructure. Collaboration Technologies and Systems (CTS), 2013 International Conference on, pp. 48-55 (2013)

[3] Duggan, J., Brodie, M.L.: Hephaestus: Data Reuse for Accelerating Scientific Discovery. (2015).

[4] Gonçalves, B. et al.: Y-DB: A system for Data-driven Hypothesis Management and Analytics. (Nov. 2014)

[5] Hey, A.J. et al. eds.: The fourth paradigm: data-intensive scientific discovery. Microsoft Research Redmond, WA (2009)

[6] Kalinichenko, L. et al.: Rule-based Multi-dialect Infrastructure for Conceptual Problem Solving over Heterogeneous Distributed Information Resources. New Trends in Databases and Information Systems. Springer International Publishing, pp. 61-68 (2013)

[7] Kalinichenko, L.A. et al.: Methods and Tools for Hypothesis-Driven Research Support: A Survey. Informatics and Application, 9 (1), pp. 28-54 (2015)

[8] Ly, D.L.,Lipson, H.: Learning Symbolic Representations of Hybrid Dynamical Systems. J. of Machine Learning Research.,13, pp. 3585-3618, Dec (2012)

[9] Porto, F. et al.: A Scientific Hypothesis Conceptual Model. Advances in Conceptual Modeling, pp. 101-110. Springer (2012)

[10] Porto, F., Schulze, B.: Data Management for eScience in Brazil. Concurrency and Computation: Practice and Experience, 25 (16), pp. 2307-2309 (2013)

[11] Robin, A., Crézé, M.: Stellar Populations in the Milky Way-A Synthetic Model. Astronomy and Astrophysics, 157, pp. 71-90 (1986)

[12] Robin, A.C. et al.: A Synthetic View on Structure and Evolution of the Milky Way. Astronomy & Astrophysics, 409 (2), pp. 523-540. (2003)

Проекты электронных библиотек

Digital library projects

Семантическое аннотирование информационных ресурсов в научной электронной библиотеке средствами таксономий

© М.Р. Когаловский[1] © С.И. Паринов[2]

[1]Институт проблем рынка РАН,
[2]Центральный экономико-математический институт РАН,
Москва

kogalov@gmail.com sparinov@gmail.com

Аннотация. Описана проблема семантического аннотирования фрагментов полных текстов публикаций, а также ссылок цитирования в публикациях научной электронной библиотеки. Предложен таксономический подход к описанию семантики аннотаций. Обсуждены основные понятия, связанные с аннотированием. Представлен ряд таксономий аннотаций, почерпнутых из литературы и опыта собственных разработок авторов. Рассмотрена реализация семантического аннотирования публикаций в научной информационной системе Соционет, которая использует также открытые данные, создаваемые средствами проекта CitEcCyr. На основе данных о содержании цитирований при просмотре публикаций в Соционет автоматически создаются аннотации внутритекстовых ссылок на используемые источники из списков литературы публикаций. Создаваемые аннотации содержат сводную информацию об источниках и статистику их цитирований.

Ключевые слова: информационный ресурс, аннотация, таксономия, цитирование, электронная библиотека, система Соционет, проект CitEcCyr.

Semantic Annotation of Information Resources by Taxonomies in Scientific Digital Library

© M.R. Kogalovsky[1] © S.I. Parinov[2]

[1]Market Economy Institute of RAS,
[2]The Central Economical and Mathematical Institute of RAS,
Moscow

kogalov@gmail.com parinov@gmail.com

Abstract. The paper discusses a semantic annotating problem with focus on full texts of research papers and citation references in publications from scientific digital library. We propose a taxonomy based approach for specifying annotation semantics. We discuss the main concepts of annotation and some annotation taxonomies taken from literature and early created by ourselves. An implementation of semantic annotating approach within the research information system Socionet is presented. This implementation is using also the open citation data created by the CitEcCyr project tools. Based on data about the content of citations while browsing the publications at Socionet automatically annotations are created for in-text references to the sources from the reference lists of publications. Generated annotations contain summary information about the sources and the statistical data about their citations.

Keywords: information resource, annotation, taxonomy, citation, digital library, Socionet system, CitEcCyr project.

1 Введение

Работая с печатным научным текстом, читатель часто делает выписки цитат или других важных для него фрагментов публикации, выделяет их в тексте, делает комментарии на полях. При работе с текстом на компьютере средствами текстовых редакторов все эти возможности также доступны. Так, версии широко распространенного текстового редактора MS Word позволяют идентифицировать фрагменты текста шрифтовым выделением или цветом, связывать с нужными фрагментами комментарии. Выделять фрагменты текста цветом и/или сопровождать их комментариями позволяют также продукты компании Adobe такие как Adobe Reader или Adobe Acrobat и некоторые другие программные

Труды XIX Международной конференции «Аналитика и управление данными в областях с интенсивным использованием данных» (DAMDID/ RCDL'2017), Москва, Россия, 10–13 октября 2017 года

средства. К сожалению, средства для таких целей не предусмотрены в стандартных веб-браузерах при просмотре страниц в формате HTML или XML, и для этого нужно использовать другие программные инструменты.

Аннотации, как результаты такой работы с текстом, читатель может создавать для собственных целей и/или для других ученых, в том числе, в процессе совместной работы по подготовке текстового документа или его экспертизе. Деятельность такого рода называется *аннотированием*. В общем случае аннотироваться могут не только тексты, но и информационные ресурсы, представленные в иных средах (графика, аудио, видео).

Аннотирование может осуществляться *в двух формах*. Первая из них заключается в дополнении к свойствам аннотируемого объекта некоторых новых атрибутов, характеризующих его дополнительные ранее не определенные свойства. Это, например, цветовое выделение фрагментов текста, тегирование музыкальных клипов, фотографий в коллекции или статей в Википедии и т. п. Вторая форма аннотирования состоит в создании нового информационного объекта, ассоциируемого с аннотируемым (целевым) объектом (субъектом аннотирования) и несущего некоторую относящуюся к нему информацию, например, комментарий, характеризующий эмоции читателя, связанные с восприятием содержания целевого объекта, или оценку его содержания, различного рода дополнения к нему и т.д. Такие вновь созданные информационные объекты, ассоциируемые с целевыми объектами, называются их *аннотациями*. В англоязычной Википедии [3] аннотацией называются «метаданные (например, комментарий, пояснение, разметка презентации), которые присоединяются к тексту, изображению или другим данным. Часто аннотации ссылаются на конкретную часть исходных данных».

В настоящее время созданы компьютерные технологии, предназначенные для аннотирования информационных объектов Веба, которые представлены в различных видах – тексты, аудио, видео и др. Однако для пользователей научных электронных библиотек и других научных информационных систем особый интерес представляет аннотирование *цифровых текстовых документов.* При этом в качестве целевых информационных объектов могут выступать не только такие документы в целом, но и их отдельные фрагменты.

В ряде развитых электронных библиотек, например, в системе Соционет [11], их информационные ресурсы включают как текстовые документы, так и различного рода связи, отражающие различные отношения между ними. В таких случаях объектами аннотирования могут быть не только фрагменты текстовых документов или документы в целом, но и связи между ними.

Аннотации сами могут представляться в виде связей между аннотатором (автором аннотации, представленным в библиотеке его персональным профилем) и аннотируемым целевым объектом. В таком случае семантика аннотации представляется семантикой этой связи.

Аннотация целевого объекта может иметь различную *семантику*, которая представляется явным или неявным образом. Если семантика представлена явным образом, то такая аннотация называется *семантической.* Соответственно, деятельность, продуктом которой являются такие аннотации, естественно называть *семантическим аннотированием.* Назначение семантической аннотации – специфицировать смысл и некоторые свойства аннотируемого ресурса.

Семантика аннотации может быть выражена *неформально*, неструктурированными метаданными, например, в виде комментария-пояснения на естественном языке, или *формально* с помощью структурированных метаданных, связывая аннотированный ресурс с некоторой семантической структурой конкретной предметной области, например, с микроформатами или с онтологией предметной области коллекции текстовых документов.

При использовании онтологии (в более простом случае – таксономии) для аннотирования используются ее классы и отношения. В случае использования онтологии для формального описания семантики аннотации аннотирование называют *онтологическим*. Могут использоваться и *комбинированные аннотации*, состоящие из формального и неформального компонентов. Например, аннотация может указывать класс таксономии, характеризующий свойство аннотируемого объекта, а также содержать текстовый комментарий на естественном языке, выполняющий аналогичную функцию или характеризующий отношение автора аннотации к целевому объекту.

Использование семантического аннотирования существенным образом обогащает восприятие информационных ресурсов пользователями, помогает интерпретировать контент аннотированных ресурсов пользователям и механизмам систем, оперирующих с ними. Оно также обеспечивает дополнительные возможности для большей полноты и точности поиска информационных ресурсов, для их анализа и обработки в больших коллекциях. На основе коллекций аннотированных научных публикаций семантические аннотации могут также использоваться для генерации различных наукометрических показателей.

Семантическое аннотирование может выполняться *вручную* экспертами, может быть *полуавтоматическим* или полностью *автоматическим,* выполняемым с помощью программных систем-аннотаторов, основанных на извлечении необходимой для этого информации из аннотируемого ресурса. Среди таких систем известны разработки, базирующиеся на наборах данных Open Linked Data (LOD) (см., например, [7]), на DBpedia [6] или Freebase [5].

В настоящей статье обсуждается подход авторов к семантическому аннотированию информационных объектов контента научных электронных библиотек

– полных текстов публикаций и их фрагментов, в частности, ссылок в тексте на используемые источники с их контекстом. Подход реализован и продолжает развиваться при участии авторов в рамках отечественной научной информационной системы Соционет [11]. Семантика аннотаций определяется средствами встроенной в систему таксономии, представленной в виде набора контролируемых словарей. Наряду с публикациями, представленными в системе, в качестве источников субъектов аннотирования используется массив описаний ссылок цитирования, автоматически генерируемый из полных текстов этих публикаций в PDF-формате [4].

Существенно отметить здесь, что специфика таксономии, используемой в нашем случае, ориентирована на описание семантики аннотаций для научных публикаций.

Остальная часть статьи организована следующим образом. В разделе 2 рассмотрен ряд проектов и публикаций, в которых предложены различные варианты таксономий аннотаций, позволяющих описывать те или иные аспекты их семантики. Особое внимание уделяется рекомендациям консорциума W3C по открытому аннотированию, также включающим один из вариантов таксономии аннотаций. В разделе 3 обсуждаются принятый подход к семантическому аннотированию в системе Соционет и его реализация. Раздел 4 посвящен обсуждению инструментария для автоматической генерации на основе полных текстов публикаций, представленных в PDF-формате, аннотаций ссылок на используемые источники. При этом аннотация включает извлеченный из полного текста контекст ссылки. Сгенерированный массив аннотаций ссылок цитирования может далее обрабатываться средствами системы Соционет. Заключение (раздел 5) подводит итоги обсуждения проблемы семантического аннотирования.

2 Таксономии аннотаций

По проблематике аннотирования вообще и семантического аннотирования, в частности, существует обширная литература, посвященная обсуждению различных подходов к аннотированию ресурсов, представленных в различных средах и относящихся к различным областям приложений, созданию стандартов в этой области, разработкам инструментария для автоматизации процесса аннотирования, подходов к семантическому аннотированию на основе различных семантических структур (систем знаний), использованию семантического аннотирования в области информационного поиска и извлечения информации из текстов, для анализа и обработки аннотированных информационных ресурсов.

Здесь мы рассмотрим несколько представленных в литературе, в том числе, разработанных авторами данной статьи подходов к описанию семантики аннотаций на основе их классификации с помощью подходящих таксономий. Иначе говоря, рассмотрим ряд походов к семантическому аннотированию на основе таксономий аннотаций, базирующих на различных их свойствах. Назовем такое аннотирование *таксономическим аннотированием*. Помимо описания семантики аннотаций использование такого подхода позволяет создавать механизмы поиска публикаций и фрагментов публикаций, адекватных потребностям пользователей, в частности, ссылок на используемые источники, а также генерировать на этой основе новые нетрадиционные наукометрические показатели.

Используемая таксономия обычно зависит от предметной области аннотируемых информационных ресурсов, целей аннотатора (эксперта или инструмента аннотирования), характера ресурсов (например, фрагменты текста или ссылки на используемые в нем источники).

Рассмотрим ряд таксономий аннотаций, предлагаемых для использования в научных электронных библиотеках. Прежде всего, обратимся к работе с привлекательным названием *"What are Semantic Annotations?"* [10]. Хотя это название обязывает авторов предложить какое-либо определение понятия *семантическая аннотация*, такого определения в явном виде в статье нет. Однако предложены общий взгляд на аннотирование и некоторая полезная систематизация сферы аннотирования. Предложения авторов статьи базируются на анализе различных подходов к аннотированию ресурсов на примере таких систем, как Semantic Wikis, Semantic Blogs, Tagging. При этом аннотирование рассматривается в общем виде как присоединение определенных данных к некоторой другой порции данных с установлением того или иного отношения между аннотированными и аннотирующими данными. Авторы различают три типа аннотаций – неформальные, формальные и онтологические. *Неформальная аннотация* представляется не на формальном языке и поэтому не является *машино-интерпретируемой* (у авторов – машиночитаемой). Напротив, *формальная аннотация* представляется на формальном языке и благодаря этому *машино-понимаема*. Однако в ней не используются термины онтологии. Наконец, *онтологическая аннотация* (которую авторы, вероятно, и понимают как семантическую) основана на использовании только терминов онтологии, и поэтому она имеет общепонятный смысл в сообществе, разделяющем эту онтологию.

В [10] предложена также общая модель аннотации, в которой предполагается, что аннотация состоит из четырех компонентов: *субъекта аннотации* – аннотируемых данных, ее *объекта* – аннотирующих данных, *предиката*, определяющего тип отношения между объектом и субъектом аннотации, и, наконец, *контекста аннотации,* характеризующего, когда и кем она создана, возможно, период времени или область пространства, где она имеет силу, и т. п. Каждый из этих компонентов может быть формальным или неформальным. Для случая аннотирования ресурсов

Веба понятия формальной и онтологической аннотации определяются более конкретно с использованием URI.

В терминах компонентов общей модели аннотации в цитируемой работе предложены заимствованные авторами из ряда публикаций *критерии* (измерения) для классификации аннотаций. Показано, какие классы аннотаций используются в каждой из анализируемых в начале статьи систем, обладающих средствами аннотирования. Использованы следующие критерии классификации аннотаций:

Ассоциация – способ, которым аннотация ассоциируется с аннотируемым ресурсом – является ли она встроенной в этот ресурс или внешней по отношению к нему и ассоциируется с ним ссылкой из ресурса;

Гранулярность субъекта аннотации – относится ли аннотация к субъекту в целом, к какому-либо его разделу или другой составной его части;

Особенность представления – аннотация относится к самому документу или к понятиям, описанным в нем либо относящимся к нему;

Повторное использование терминологии – использует ли аннотация собственную терминологию или термины из существующих онтологий и тем самым интероперабельна и понятна для других;

Тип объекта – является ли объект аннотации литеральным или текстовым, структурным или онтологическим;

Контекст – контекст аннотации: когда, кем она создана, в какой сфере, какой срок ее действительности и т. п.

Предложенная классификация аннотаций, хотя и не полна, по нашему мнению, полезна для описания не семантики аннотаций, создаваемых в той или иной электронной библиотеке, скорее, функциональных возможностей используемого в конкретной системе подхода к аннотированию и/или конкретных инструментов семантического аннотирования, а также для сопоставления функциональности различных таких подходов/инструментов.

Значимый вклад в создание технологий и инструментария интероперабельного аннотирования, основанного на формальном языке представления аннотаций, вносит деятельность Группы по открытому аннотированию (Open Annotation Group или кратко OAG), функционирующей в последние годы в рамках консорциума W3C. Эта группа разрабатывает спецификации стандарта онтологии (в терминологии группы – *модели данных*), описываемой на языке RDF, и протокола для открытого интероперабельного аннотирования цифровых документов – текстов, графических изображений, аудио, таблиц и других ресурсов, а также их фрагментов.

В настоящее время предложенные группой спецификации приобрели статус рекомендации консорциума [15–17] и рассматриваются как средство для Семантического Веба, хотя некоторые их элементы могут иметь и более широкое применение.

В спецификациях OAG предложена онтология аннотирования, формально определяющая различные виды аннотаций: комментарии, аннотации сущностей (или как теперь принято говорить, вещей), заметок, примеров, опечаток и т. п.

Таблица 1

№/№ п.п.	Мотивация	Пояснение
1.	Оценивание	Аннотация служит для оценки целевого ресурса.
2.	Установка закладки	Аннотация отмечает некоторое указанное ее автором место в тексте целевого ресурса.
3.	Классифицирование	Аннотация используется для классификации целевого ресурса.
4.	Комментирование	Аннотация представляет собой комментарий, относящийся к целевому ресурсу.
5.	Описание	Аннотация служит для описания свойств целевого ресурса.
6.	Редактирование	Аннотация указывает необходимость редактирования целевого ресурса, например, с тем чтобы устранить опечатку.
7.	Выделение маркером	Аннотация указывает намерение ее автора выделить цветом целевой ресурс или его фрагмент для того, чтобы по какой-то причине обратить на него внимание.
8.	Идентификация	Аннотация служит для придания индивидуальности целевому ресурсу путем ассоциирования с ним какого-либо уникального идентификатора, например, URI.
9.	Связывание	Аннотация определяет связь с некоторым ресурсом, имеющим отношение к целевому.
10.	Модерирование	Аннотация служит для указания ценности или качества целевого ресурса, например, для модерирования дискуссий и обсуждений.
11.	Запрашивание	Аннотация содержит вопрос о целевом ресурсе.
12.	Ответ	В аннотации приводится отклик на целевой ресурс.
13.	Создание пометы	Аннотация содержит помету для целевого ресурса.

В контексте данной статьи представляет интерес используемый в онтологии контролируемый словарь мотивов, которыми руководствуется создатель аннотаций. Этот словарь, по существу, может рассматриваться как таксономия мотивов аннотирования, позволяющая явным образом специфицировать их семантику. Классы словаря мотивов аннотирования приведены в таблице 1.

Частным случаем связей между текстовыми документами в электронной библиотеке являются связи цитирования, представляемые в виде ссылок на используемые или упоминаемые в данной публикации источники вместе с контекстами этих ссылок. Такие ссылки, как и другие связи, могут стать субъектами аннотирования наряду с текстовыми документами или их фрагментами. С позиций аннотирования целесообразно различать разные виды ссылок на использованные источники: ссылки с контекстом – цитатой из цитируемого источника, ссылки с иным контекстом и, наконец, ссылки на источники, указанные в списке литературы, но с отсутствующими на них ссылками в тексте.

Для семантического аннотирования ссылок цитирования также могут использоваться таксономии ссылок. В ряде публикаций содержатся предложения подходящих для этого таксономий. Например, в работе [8], посвященной анализу категоризации влияния цитируемых источников на цитирующие публикации, предлагается классификация ссылок цитирования в трех измерениях: *функция* (*Function*), *полярность* (*Polarity*) и *влияние* (*Impact*). Для каждого из этих измерений предложен свой набор классов. Измерению *функция* соответствуют классы, указывающие, что цитируемый источник полезен (*Useful*), отражает противоположную точку зрения (*Contrast*), обладает недостатками (*Weakness)*, вносит поправки *(Correct*), уклоняется (*Hedges*), выражает благодарность (*Acknowledge)*, подтверждение (*Corroboration)*, полемизирует *(Debate*). Для измерения *полярности* предлагаются следующие классы: позитивная (*Positive*), негативная (*Negative*) и нейтральная (*Neutral*). Наконец, для измерения *влияния* предложены такие классы: негативное (*Negative)*, незначительное (*Perfunctory*) и существенное (*Significant)*.

В работе [14] также предложена классификация ссылок цитирования. Используются иные критерии по сравнению с рассмотренными выше. Ссылки классифицируются *по месту в тексте* и ранжируются таким образом, что выше их ранг в разделе с результатами, ниже в обзоре литературы, *по количеству вхождений*, а также *по стилю.* В качестве места в тексте рассматриваются его разделы: абстракт, введение, обзор литературы, методология, результаты/обсуждение, заключение. Возможные варианты стиля: неконкретное упоминание (not specially), конкретное и интерпретирующее упоминание, прямая цитата.

В используемых в настоящее время описаниях ссылок цитирования отсутствуют атрибуты, которые бы позволили отобразить их классификацию по критериям значимости (место в тексте), интенсивности (частотности) и по стилю, предложенным в рассматриваемой статье. Чтобы их специфицировать, достаточно ввести в таксономию два контролируемых словаря:

- *Словарь мер* (или интенсивностей): высокая, средняя, низкая. Его следует использовать для характеристики значимости ссылки (в зависимости от места в тексте) и оценки частотности.
- *Словарь стилей* (характер контекста): прямая цитата, неконкретное упоминание источника, упоминание с пояснением, ссылка без контекста (для случая ссылки в списке литературы, не упоминаемой в тексте).

На основе приведенной классификации ссылок цитирования с помощью указанных контролируемых словарей могут генерироваться новые наукометрические показатели, например, следующие: *количество ссылок высокой* (а также средней/низкой) значимости на данную работу, *количество ссылок с высокой* (а также со средней/низкой интенсивностью), *количество ссылок с прямым цитированием* (а также с интерпретацией в контексте/с неконкретным контекстом/без контекста).

Необходимо упомянуть также онтологию ссылок цитирования C4O (the Citation Counting and Context Characterization Ontology) [12], представляющую собой составную часть модульного комплекса онтологий SPAR [13], некоторые элементы которых ранее уже были использованы в таксономии системы Соционет. Онтология C4O включает важные для нашей работы классы отношений между источниками из списков литературы и ссылками на них в текстах публикаций. Эти вопросы обсуждаются ниже в разд. 4.

Таксономический подход для описания семантики аннотаций используется и в системе Соционет. В этой системе поддерживается встроенная таксономия [1], используемая для классификации и тем самым для описания семантики связей между информационными объектами контента системы. Некоторые контролируемые словари, составляющие эту таксономию, используются и для семантического аннотирования. В частности, для этой цели можно использовать оценочный контролируемый словарь. Этот словарь может использоваться не только для аннотирования полного текста публикации и ее фрагментов, но также и ссылок на использованные источники в тексте публикации, а также в послестатейном списке литературы. Во всех указанных случаях, кроме последнего, аннотирование может осуществлять любой авторизованный пользователь системы, в последнем случае – только автор данной публикации. Оценочный контролируемый словарь включает, в частности, следующие классы: наилучшая, наиболее релевантная работа по обсуждаемой в ней теме;

новаторская работа (результат); интересная работа (результат); оценивается позитивно; оценивается негативно; основывается на заблуждении; возможно, является плагиатом.

Встроенная в систему Соционет таксономия может легко расширяться путем дополнения новых контролируемых словарей, позволяющих описывать новые аспекты семантики аннотаций. Обсуждается дополнение таксономии рядом новых словарей. Для аннотирования фрагментов авторефератов диссертаций и полных текстов диссертаций полезен словарь, позволяющий идентифицировать в текстах этих документов важные для их оценки оппонентами фрагменты, содержащие аргументацию соответствия диссертации требованиям ВАК. Словарь включает классы: *актуальность, новизна, достоверность, практическая ценность, теоретическая ценность.* Полезен также контролируемый словарь, позволяющий специфицировать *статус* аннотируемых фрагментов полного текста публикации: *аксиома*, *доказанное утверждение* (теорема), *цитата из используемого источника*, *фактография*, *результат исследования, постановка задачи.* Может быть также расширен оценочный словарь дополнительным включением в него следующих дополнительных классов: *актуальная тема исследования, актуальный результат, оригинальный результат, уже известный в науке результат, новый научный результат, фундаментальный результат, обоснованное утверждение, необоснованное утверждение, вода, раскавыченная цитата.*

Рассмотренные таксономии показывают, что их конкретные варианты следует использовать в соответствии с характером аннотируемых ресурсов и целями аннотатора.

3 Семантическое аннотирование в Соционет

В системе Соционет обеспечиваются возможности открытого семантического аннотирования. Важно отметить, что они реализуются с использованием тех же средств, которые уже имелись в системе для создания, поддержки и использования семантических связей между информационными объектами ее контента. Использовать возможности семантического аннотирования может зарегистрированный и авторизовавшийся пользователь, поскольку предусматривается фиксация авторства созданных аннотаций.

В Соционет поддерживаются информационные объекты – научные публикации, научные отчеты и научные произведения других видов, и семантические связи между ними [2]. Семантика связей определяется с помощью встроенной в систему таксономии, состоящей из нескольких контролируемых словарей. Эта таксономия подробно рассмотрена в работе [1] и кратко обсуждена вместе с некоторыми возможными ее расширениями в предыдущем разделе. Классы некоторых контролируемых словарей таксономии используются для описания семантики аннотаций. Это естественный подход, поскольку аннотации представляются в системе в виде семантических связей.

С точки зрения общей модели аннотаций, предложенной в [10], модель аннотаций, используемую в Соционет, можно назвать *комбинированной* – объект аннотации включает формальный и неформальный компоненты. Формальный компонент – это структурированные метаданные, указывающие один из классов подходящего контролируемого словаря встроенной в систему таксономии, определяющий семантику аннотации. Неформальный компонент, называемый в описании аннотации комментарием, – это неструктурированные метаданные, представленные в виде текста на естественном языке.

Субъектами аннотирования в Соционет могут быть полные тексты представленных в системе публикаций, фрагменты их абстрактов, а также фрагменты полных текстов. Кроме того, аннотироваться могут также и связи цитирования одних публикаций в других. Связи этого вида – это ссылки на источники из послестатейного списка литературы, а также сами библиографические описания использованных источников в этих списках.

Наряду со связями цитирования, выделяемыми пользователем-аннотатором в «ручном режиме», субъекты аннотирования такого рода могут порождаться в автоматическом режиме средствами анализа полных текстов публикаций, представленных в контенте системы в pdf-формате. Эта техника и ее возможности обсуждаются в следующем разделе.

Соционет является мультипользовательской системой, и поэтому для одного субъекта аннотирования может быть создано несколько аннотаций одним или разными пользователями системы. Аннотации представляются в Соционет в виде классифицированных связей «персона – субъект», и их описания включают идентификацию персоны-автора аннотации, идентификацию субъекта аннотации, класс выбранного аннотатором контролируемого словаря таксономии, а также текстовый комментарий.

Функциональные возможности системы Соционет позволяют использовать ее как платформу для виртуальной коммуникационной среды научного сообщества пользователей системы [9]. Эти возможности основаны на реакциях авторов публикаций, представленных в системе, на появлении семантических связей этих публикаций с публикациями других авторов либо оценочных связей, касающихся этих публикаций. Такая реакция состоит в создании новой связи профиля ее автора со связью, на появление которой он реагирует. Поскольку аннотации представляются в виде семантических связей, указанные возможности могут быть применены и к ним. Поэтому, хотя такая возможность пока еще не полностью реализована в

Соционет, создание аннотаций потенциально может быть вовлечено в возникающие в такой среде процессы коммуникаций, отображающие дискуссии относительно создаваемых аннотаций.

Формальные компоненты объектов аннотаций – структурированные метаданные – могут использоваться в критериях поиска аннотаций, интересующих пользователя классов, а также для генерации ряда новых наукометрических показателей наряду с другими, формируемыми сервисами системы. Для возможности генерации новых наукометрических показателей в описания создаваемых связей должны быть перенесены классификационные атрибуты цитирования. Должны быть также созданы в Соционет соответствующие сервисы, которые будут генерировать и показывать полученные показатели на странице метаданных (описателя) публикации, как это реализовано сегодня для других показателей в системе. Этими новыми показателями могут быть, например, следующие: *количество ссылок высокой* (а также средней/низкой) значимости на данную работу, *количество ссылок с высокой* (а также со средней/низкой интенсивностью), *количество ссылок с прямым цитированием* (а также с интерпретацией в контексте/с неконкретным контекстом/без контекста).

4 Генерация описаний ссылок цитирования и их визуализация в Соционет

Интересные перспективы для развития семантического аннотирования в системах, подобных Соционет, открывают новые подходы и технологии, создаваемые для поддержки анализа содержания цитирований. Общая концепция анализа содержания цитирований представлена в [14]. Описание создаваемых технологий, применение которых обсуждается в данной статье ниже, доступно в [4]. Основная новизна этих подходов связана с извлечением из научных публикаций более широкого по сравнению с традиционным подходом набора данных, связанных со ссылками цитирования, включая окружающий их контекст. Кроме того, создаются новые возможности визуализации этих данных, которые позволяют накладывать результаты анализа содержания цитирований поверх текста pdf-документов в виде программным образом генерируемых аннотаций.

Проект CitEcCyr (https://github.com/citeccyr), реализуемый с участием одного из авторов данной статьи в Российской академии народного хозяйства и государственной службы при Президенте РФ (РАНХиГС) с 2016 г., предусматривает разработку средств извлечения из русскоязычных научных публикаций, доступных в виде pdf-документов, расширенного набора сведений о цитированиях. На основе этих данных предполагается разработка новых наукометрических показателей, включая некоторые дополнительные данные о научной результативности. Предполагается учитывать количество ссылок в тексте публикации на источники из списка литературы, отделять источники без ссылок на них в тексте публикации. Кроме того, имеется в виду обрабатывать контекст вокруг ссылок для классификации содержания цитирований источников, а также ранжировать ссылки на источники по месту их в структуре статьи (например, ранг выше в разделе с результатами, ранг ниже в разделе обзор литературы) и др.

Источником публикаций для обработки средствами проекта является система Соционет. Первые результаты извлечения данных о цитированиях, полученные на основе публикаций архива НЭИКОН (https://socionet.ru/collection.xml?h=spz:neicon), доступны для ознакомления и тестирования по адресу http://no-xml.socionet.ru/~cyrcitec/citmap/spz/neicon/.

Средствами обсуждаемого проекта создаются новые данные о цитированиях. Рассмотрим их особенности, а также их визуализацию в Соционет на примере одной из научных публикаций гуманитарного профиля, доступной в виде pdf-документа по адресу http://nevolin.socionet.ru/files/2014_Nevolin_ rfbr.pdf.

На Рис. 1 приведен фрагмент этой публикации, в котором на экране компьютера ссылки на источники из списка литературы выделяются желтым цветом. К этим выделенным фрагментам текста публикации программным образом созданы аннотации. Кликая на выделенные цветом ссылки, пользователь получает различную дополнительную информацию.

Чтобы это стало возможным, создаваемые с помощью программного обеспечения проекта CitEcCyr данные о цитированиях преобразуются в соответствии с моделью данных веб аннотаций [15]. Затем эти данные интегрируются в среду системы Соционет в виде семантических связей, что является обычным для представления аннотаций в системе. Рассмотрим, какова общая схема получения этих данных о цитированиях и как они в данном случае используются.

На первом этапе выполняется конвертация бинарных pdf-документов в текстовый вид, который допускает анализ и извлечение необходимых данных о цитированиях. В проекте CitEcCyr разработана программа конвертации PDF-STREAM (https:// github.com/citeccyr/pdf-stream-cli), которая преобразует содержание pdf-документов в формат JSON. Пример данных, получаемых для указанной выше публикации, доступен по адресу http://no-xml.socionet.ru/citmap/convertedPDF/2014_Nevolin_ rfbr.json.

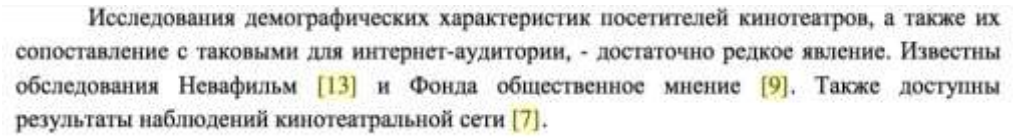

Исследования демографических характеристик посетителей кинотеатров, а также их сопоставление с таковыми для интернет-аудитории, - достаточно редкое явление. Известны обследования Невафильм [13] и Фонда общественное мнение [9]. Также доступны результаты наблюдений кинотеатральной сети [7].

Рисунок 1 Пример фрагмента публикации с выделенными аннотированными ссылками на цитируемые источники, одновременно служащими указателями на аннотации

На следующем этапе работает программа, также созданная в проекте CitEcCyr, которая создает XML-записи, содержащие, в том числе, сведения о ссылках на источники из списка литературы публикации. Для упомянутой выше публикации на основе ее JSON-версии ниже приведен пример XML-записи, содержащей извлеченные данные для двух ссылок (в тегах <intextref>) на один и тот же источник, который имеет в списке литературы порядковый номер 7.

```
<intextref>
	<Reference>7</Reference>
	<Exact>[7]</Exact>
	<Start>6125</Start>
	<End>6128</End>
	<Prefix>сом –имеются данные, что реклама и продажи в баре составляют, соответственно,
20-25% и 20-30% выручки кинотеатров</Prefix>
	<Suffix>.Итак, характеристики аудитории представляют коммерческую ценность для отрасли
и научный интерес для исследователей, н</Suffix>
</intextref>

<intextref>
	<Reference>7</Reference>
	<Exact>[7]</Exact>
	<Start>10119</Start>
	<End>10122</End>
	<Prefix>обследования Невафильм[13] и Фонда общественное мнение[9].
Также доступны результаты наблюдений кинотеатральной сети</Prefix>
	<Suffix>. Согласно данным Невафильм (см. Таблицу 2), профили аудиторий-посетителей
кинотеатров и интернет-пользователей, -з</Suffix>
</intextref>
```

Эти данные включают:

- номер источника в списке литературы, тег <Reference>, в примерах выше он содержит номер 7;
- вид ссылки на соответствующий источник, тег <Exact>, в примерах это - [7];
- текстовые координаты ссылки в тегах <Start> и <End>, которые содержат порядковые номера от начала текста документа первого и последнего символа строчки, содержащейся в теге <Exact>;
- контекст вокруг ссылки в тегах <Prefix> и <Suffix>, который в данном случае включает по 200 символов слева и справа от ссылки, содержащейся в теге <Exact>.

В частности, второй блок данных в теге <intextref> из приведенной выше XML-записи отображен на Рис. 1 как аннотация к ссылке на источник номер 7.

Кроме этого, из JSON-версии pdf-документов извлекаются данные о содержании списка литературы публикаций, которые иллюстрируются следующей XML-записью:

```
<reference>
	num="7"
	start="20952"
	end="21140"
	author="Гладких Михайлина"
	title="Кронверк Синема сколько стоит билет в кино"
	year="2011"
	handle="spz:cyberleninka:33099:16516633">
	<from_pdf>Гладких И.В., Михайлина А.П. «Кронверк Синема»: сколько стоит билет в кино?
(учебный кейс) / Вестник Санкт Петербургского университета. Серия 8: Менеджмент. 2011. No3.
с.145 159.</from_pdf>
</reference>
```

Эти данные содержат:

- атрибут num – номер источника в списке литературы, в приведенном выше примере он - номер 7;
- атрибуты start и end – текстовые координаты данных источника, которые содержат порядковые номера от начала текста документа первого и последнего символа строчки, содержащейся в теге <from_pdf>;
- атрибуты author, title и year, выделенные из данных тега <from_pdf> и используемые для поиска в Соционет публикации, которая указана в данных этого источника;

- атрибут handle – содержит уникальный код публикации, соответствующей данным этого источника, если она есть в Соционет;
- тег <from_pdf> – содержит «сырые» данные источника, которые извлечены из JSON-версии публикации.

Полный набор данных о содержании цитирований для научной статьи, к которому относятся приведенные выше примеры, доступен по адресу http://no-xml.socionet.ru/citmap/outputs/repec:rus:pgfhxz:wp9.xml.

Описанные выше данные о содержании цитирований допускают различные варианты их использования в научных информационных системах, подобных Соционет. Поскольку данные о ссылках на цитируемые источники включают их текстовые координаты, то возможно программное создание аннотаций, которые при просмотре соответствующих публикаций в Соционет выглядят визуально привязанными к тексту ссылок на источники. На Рис. 2 приведен пример визуализации данных о цитированиях в виде аннотаций ссылок на цитируемые источники, выделяемых цветом на экране компьютера. В частности, для ссылки на 7-й источник раскрыта ее аннотация (справа), которая в текущей версии содержит данные о соответствующем источнике и, если есть, ссылку на него в Соционет, а также статистику о количестве цитирований данного источника в этой публикации.

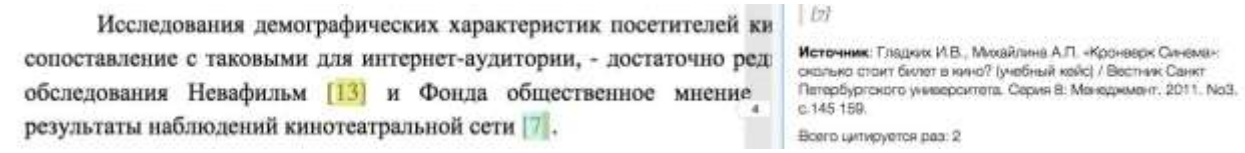


Рисунок 2 Пример программно-сгенерированной аннотации для ссылки на источник

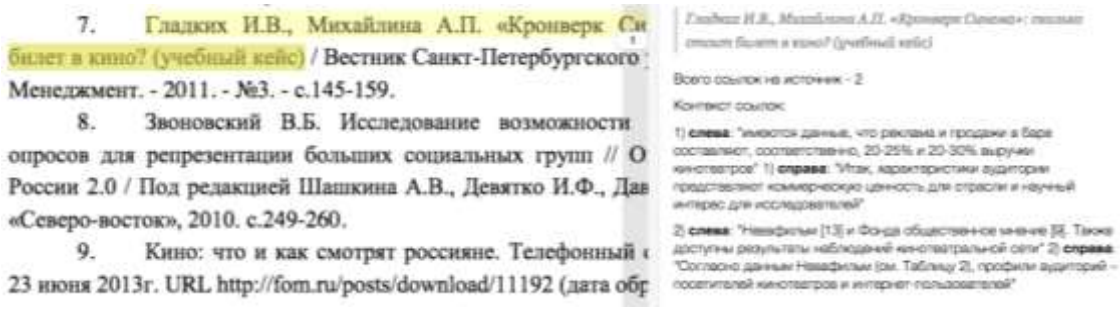


Рисунок 3 Пример программно сгенерированной аннотации для библиографического описания источника

В будущем планируется также приводить статистику обо всех цитированиях данного источника в контенте Соционет.

Похожим образом могут быть построены аннотации поверх данных о библиографических описаниях источников в списке литературы публикации. На Рис. 3 приведен пример аннотации в списке литературы, которая «наложена» поверх публикации с номером 7. Справа в текстовом блоке видно содержание этой аннотации.

Аннотация на Рис. 3 содержит сведения об общем количестве упоминаний (цитирований) данного источника в тексте публикации, а также контекст (по 200 символов справа и слева) для каждого такого случая.

Рассмотренные технологии аннотирования позволяют в нужных местах текста публикаций компактно предоставлять пользователям Соционет различную дополнительную информацию. Эта информация, как это представлено выше, может содержать обобщенные данные о содержании цитирований. Уже имеющиеся в Соционет сервисы для авторов публикаций для «ручного» семантического «раскрашивания» связей, в данном случае, позволяют им как уточнять программно-сгенерированную семантику аннотаций, так и добавлять к аннотациям новые семантические атрибуты.

5 Заключение

Современные научные информационные системы, к числу которых относится и система Соционет, начинают предлагать своим пользователям различные возможности для семантического аннотирования контента. Сравнительно новыми возможностями является доступное в Соционет «ручное» аннотирование полных текстов научных статей, представленных в виде pdf-документов, и их фрагментов. В дополнение к этому в Соционет разрабатываются средства программной генерации аннотаций для ссылок цитирования, которые являются важным элементом научных публикаций и академической культуры. Данный подход позволяет через аннотации, привязанные к определенным фрагментам pdf-документов, показать читателю разнообразную наукометрическую информацию, включая сводные сведения о том, сколько раз цитируются источники из списка литературы в данной публикации, а также и во всех других публикациях, имеющихся в системе Соционет.

Благодарности

Реализация методов аннотирования в системе Соционет выполнена в рамках работ по гранту РФФИ, проект 15-07-01294. Разработка подхода для извлечения данных о содержании цитирований, в том числе, для целей суперкомпьютерного моделирования взаимодействий между агентами и со средой научного сообщества, были получены С.И. Париновым в рамках работ по гранту РНФ, проект 14-18-01968.

Литература

[1] Когаловский, М.Р., Паринов, С.И.: Таксономия семантических связей информационных объектов контента научной электронной библиотеки. НТИ. Серия 2. Информационные процессы и системы, 9, сс. 15-23 (2015)

[2] Паринов, С.И., Когаловский, М.Р.: Технология семантического структурирования контента научных электронных библиотек. RCDL 2011, pp. 197-206 (2011)

[3] Annotation. Wikipedia. https://en.wikipedia.org/wiki/Annotation

[4] Barrueco, J.M., Krichel, T., Parinov, S., Lypunov,V., Medvedeva, O., Sergeeva, V.: Towards Open Data for Citation Content Analysis. Submitted to DAMDID/RCDL-2017

[5] Bennet, P.N., Gabrilovich, E., Kamps, J., Karlgren, J.: Sixth Workshop on Exploiting Semantic Annotations in Information Retrieval (ESAIR'13). CICM'13, pp. 2543-2544 (2013)

[6] DBpedia. Википедия. https://ru.wikipedia.org/wiki/DBpedia

[7] Gagnon, M., Zouaq, A., Jean-Louis, L.: Can we use Linked Data Semantic Annotators for the Extraction of Domain-Relevant Expression. WWW 2013 Companion, pp. 1249-1246 (2013)

[8] Hernández-Alvarez, M., Gómez Soriano, J.M., Martinez-Barco, P.: Citation Function, Polarity and Influence Classification. doi: 10.1017/S1351324916000346 (2017)

[9] Kogalovsky, M.R., Parinov, S.I.: Scholarly Communications in a Semantically Enrichable Research Information System with Embedded Taxonomy of Scientific Relationships. In: Klinov, P. and Mouromtsev, D. (eds.): Knowledge Engineering and Semantic Web. 6th Int. Conf. KESW 2015. The Communications in Computer and Information Science series, 518. Springer, pp. 87-101 (2015)

[10] Oren, E., Hinnerk Moller, K., Scerri, S., Handschuh, S., Sintek, M. What are Semantic Annotations? (2006) http://www.siegfried-handschuh.net/ pub/2006/whatissemannot2006.pdf

[11] Parinov, S., Lyapunov, V., Puzyrev, R., Kogalovsky, M.: Semantically Enrichable Research Information System SocioNet. In: Klinov, P. and Mouromtsev, D. (eds.): Knowledge Engineering and Semantic Web. 6th Int. Conf. KESW 2015. The Communications in Computer and Information Science series, 518. Springer, pp. 147-157 (2015)

[12] Shotton, D.: C40, the Citation Counting and Context Characterization Ontology. Version 1.1.1, 11/05/2013. http://purl.org/spar/c4o

[13] SPAR Ontologies. Describing Publishing Domain. http://purl.org/spar/

[14] Zhang, G., Ding, Y., Milojević, S.: Citation Content Analysis (CCA): A Framework for Syntactic and Semantic Analysis of Citation Content. arXiv:1211.6321 (2012)

[15] Web Annotation Data Model. W3C Recommendation 23 February 2017. https://www.w3.org/TR/2017/REC-annotation-model-20170223/

[16] Web Annotation Protocol. W3C Recommendation 23 February 2017. http://www.w3.org/TR/annotation-protocol/

[17] Web Annotation Vocabulary. W3C Recommendation 23 February 2017. http://www.w3.org/TR/annotation-vocab

Принципы создания многоязычной электронной библиотеки для крупного информационного центра

© В.Н. Захаров © Ю.В. Никитин © Ал-др А. Хорошилов © Ал-ей А. Хорошилов

Федеральный исследовательский центр «Информатика и управление» РАН, Москва, Россия

vzakharov@ipiran.ru yuri.v.nikitin@gmail.com khoroshilov@mail.ru a.a.horoshilov@mail.ru

Аннотация. Описан подход к созданию многоязычной электронной библиотеки для крупного информационного центра. Показано, как организовать процесс формализации документов на разных языках таким образом, чтобы поиск был максимально эффективен и позволял пользователю получать результаты независимо от языка запроса и документов, содержащихся в базе данных. Исследование эффективности предложенного подхода показало достаточно высокие результаты, позволяющие применять его в промышленных информационных системах.

Ключевые слова: многоязычная электронная библиотека, многоязычный поиск, информационный поиск, автоматизированная обработка текстов, формализованное описание текста, смысловая структура, лингвистическое программное обеспечение, декларативные средства.

The Principles of Creating a Multilingual Electronic Library for a Large Information Center

© V.N. Zakharov © Yu.V. Nikitin © Al-dr A. Khoroshilov © Al-ey A. Khoroshilov

Federal Research Center Computer Science and Control of the Russian Academy of Sciences, Moscow, Russia

vzakharov@ipiran.ru yuri.v.nikitin@gmail.com khoroshilov@mail.ru a.a.horoshilov@mail.ru

Abstract. This paper describes the approach to creating a multilingual electronic library for a large information center. The authors show how to organize the process of formalizing documents in different languages in such a way that the search is most effective and allows the user to receive results regardless of the query language and documents contained in the database. The study of the effectiveness of the proposed approach has shown quite good results, allowing it to be used in industrial information systems.

Keywords: multilingual electronic library, multilingual search, information retrieval, automated text processing, formal description of text, semantic structure, linguistic software, declarative means.

1 Введение

В нашей стране в настоящее время функционирует множество организаций, каждый день имеющих дело с огромным объемом документов. Многие из этих организаций из-за специфики своей деятельности получают и обрабатывают документы на нескольких языках. К таким организациям можно отнести, например, предприятия авиационно-косми-ческой отрасли, для которых стоит важнейшая задача соответствия международным стандартам; всевозможные научные организации, для которых жизненно необходимо быть в курсе последних исследований и разработок; организации, обеспечивающие безопасность государства, для получения актуальной политической и технической информации и т. д. Соответственно, для решения различных задач дальнейшего эффективного использования получаемых постоянно документов необходима их предварительная автоматическая обработка, позволяющая свести к минимуму трудозатраты обслуживающего персонала. В настоящее время множество информационных систем имеет достаточно полный функционал работы с русскоязычными текстами, но, к сожалению, все эти системы имеют довольно скромные возможности при работе с разноязычными массивами документов, а задача сравнения текстов, выявления документов-дубликатов и заимствований в отечественных информационных системах в настоящий момент решена только для документов, написанных на одном языке. В то же время потребность в таких системах достаточно велика, и задача требует скорейшего решения.

Труды XIX Международной конференции «Аналитика и управление данными в областях с интенсивным использованием данных» (DAMDID/ RCDL'2017), Москва, Россия, 10–13 октября 2017 года

2 Существующие подходы к организации электронных библиотек

Задача хранения и организации доступа к большим коллекциям документов стоит уже достаточно давно. За это время было разработано множество решений, которые в разной степени удовлетворяют требованиям, предъявляемым современными пользователями. Далее приведем некоторые данные о развитии такого программного обеспечения в настоящее время.

В работе [1] авторы провели серьезное сравнение свободно распространяемых технологий для организации электронных библиотек, существующих в настоящее время. Были протестированы системы OJS, ePubTK, DPubS, GAPWorks, Ambra, e-Journal. Сделан вывод, что практически все решения поддерживают общепринятые стандарты в области интеграции и обмена данными и имеют широкие возможности по генерации различных метаданных в зависимости от потребностей пользователя. Но, к сожалению, большинство из рассматриваемых продуктов более не развивается. Понятно, что такие системы позволяют решать стандартный набор задач и используются для небольших электронных библиотек.

При росте объемов документов становится важно решить задачу повышения эффективности поиска. Для этого многие ученые разрабатывают новые механизмы, одним из которых стал семантический поиск. В работе [2] авторы предлагают новый метод поиска, основанный на использовании модели S-тег. Особенностью данного метода является то, что индексируется не весь текст, а только его значимые части в зависимости от задачи, при этом за счет изменения размера значимой части можно контролировать точность и полноту.

Другим подходом к семантическому поиску, о котором сейчас пишет все большее число авторов, является использование онтологических моделей [3]. Основной идеей данного подхода является использование онтологий предметных областей для аннотирования содержания электронных ресурсов. Авторы работы [4] дополнили онтологический подход добавлением новых операций над онтологиями – проекции и масштабирования – и описали модель их применения для задач информационного поиска.

Еще одним направлением развития поиска в электронных библиотеках является многоязычный поиск. К сожалению, в настоящее время работ по этой тематике не так много. Одно из таких решений было описано в работе [5]. В ней представлено решение задачи двуязычного поиска с помощью тезауруса для двух языков (русского и английского). Похожего мнения придерживаются и многие иностранные исследователи, в том числе, например, в работе [6]. Несколько иной подход предложен в [7]: для решения задачи многоязычного поиска использован инструментарий систем автоматического перевода текстов.

3 Организация многоязычной электронной библиотеки для крупного информационного центра

3.1 Архитектура многоязычной электронной библиотеки

Проанализировав подходы и решения, имеющиеся на сегодня в области разработки современных электронных библиотек, авторами был составлен список требований, которым должна удовлетворять система, функционирующая в крупном информационном центре:

- обеспечение модульной архитектуры с возможностью быстрого включения в систему новых модулей;
- использование средств СУБД, позволяющих максимально эффективно организовать процесс доступа к данным;
- обеспечение возможности оперативного пополнения декларативных средств системы;
- обеспечение максимальной простоты добавления новых языков в систему;
- обеспечение распределенной массово-параллельной лингвистической и статистической обработки загружаемых данных;
- обеспечение масштабируемости на множество узлов обработки без деградации инфраструктуры обработки данных;
- обеспечение всех этапов лингвистической обработки, включающей этапы графематического, морфологического, семантико-синтаксического, концептуального и дистрибутивно-статистического анализа [12];
- обеспечение эффективного многоязычного поиска;
- обеспечение эффективного сравнения смыслового содержания документов, в том числе поиска заимствований и документов-дубликатов [13-15];
- обеспечение поддержки общепринятых стандартов в области интеграции и обмена данными;
- создание наиболее полной и удобной структуры метаданных для хранимых в базе документов;
- обеспечение удобного пользовательского интерфейса, максимально упрощающего доступ пользователя ко всему функционалу электронной библиотеки.

На Рис. 1 представлена предлагаемая авторами архитектурная схема многоязычной электронной библиотеки для крупного информационного центра.

Рисунок 1 Архитектурная схема многоязычной электронной библиотеки для крупного информационного центра

3.2 Процесс формализации документов в многоязычной электронной библиотеке

Основной задачей при выполнении формализации документа является представление смысловой структуры текста в структурированном виде. По мнению авторов, формализованное представление текстового содержания документа должно включать:

- библиографические реквизиты (например, информационный источник, рубрика, автор, наименование и дата публикации и т. п.);
- аннотацию или реферат документа;
- список ключевых выражений;
- список значимых объектов (персоны, организации, территории, наименования товаров, географические объекты, бренды, и т. д.);

При этом для создания многоязычной системы данная информация должна содержаться на всех поддерживаемых языках. Также каждому документу должна соответствовать следующая информация:

- содержащиеся в документе формулы, параметры с их числовыми значениями и т. д.;
- классификация документа по смысловому содержанию – отнесение его к той или иной рубрике и кластеризация [11] (группировка) текстов публикаций по темам;
- ссылки на связанные документы (цитаты, заимствования, документы-дубликаты, близкие по смыслу документы) [8–10].

3.3 Организация многоязычного поиска

В ходе исследования авторами была разработана двухступенчатая процедура поиска, которая может быть использована для поиска в многоязычном массиве информации. На первом этапе запрос был преобразован в его унифицированное семантическое представление, на втором этапе производился поиск в базе данных стандартными средствами. Рассмотрим каждый из этапов подробнее.

3.3.1 Метод трансформации поискового запроса в его унифицированное семантическое представление

Разработанный авторами метод трансформации поискового запроса в его унифицированное семантическое представление основан на использовании многоязычного словаря унифицированных формализованных представлений наименований понятий [16]. В данном исследовании словарь был сформирован для трех языков (русского, английского и немецкого), но в этот словарь могут быть добавлены эквиваленты на других языках при наличии переводных словарей схожих объемов. Также для работы метода необходимы процедуры морфологического, семантико-синтаксического и концептуального анализа для каждого языка, который содержится в словаре унифицированных формализованных представлений наименований понятий. При выполнении этих условий трансформация поискового запроса сводится к следующему алгоритму (Алгоритм 1):

Шаг 1. Определяется язык обрабатываемого запроса.

Шаг 2. С помощью процедуры концептуального анализа (для выявленного языка) определяется совокупность значимых наименований понятий с указанием местоположений этих понятий в тексте запроса.

Шаг 3. Каждое наименование понятия запроса приводится к нормальной форме с помощью процедуры автоматической пословной нормализации.

Шаг 4. Каждое нормализованное наименование понятия ищется в многоязычном словаре унифицированных формализованных представлений наименований понятий, после чего ему присваивается номер из этого словаря. Пример словаря приведен в Таблице 1.

Таблица 1 Фрагмент многоязычного словаря унифицированных формализованных представлений наименований понятий

№ n/n	Значения на русском языке	Эквиваленты на английском языке	Эквиваленты на немецком языке
…	…	…	
816437	нефтехранилище / нефтесклад / хранилище	oil reservoir / oil storage / petroleum storage / tank farm	öllager / erdöllager / tanklager
816438	нефть / каустобиолит / петролеум / черный золото	mineral oil / naphtha / oil / petrol / petroleum / rock-oil	öl / caustobiolith / petroleum / petrol
816439	нефтяник / нефтедобытчик	oilman / oil-industry worker	ölproduzent / ölhändler
…	…	…	

Схема работы данного алгоритма отображена на Рис. 2.

Рисунок 2 Схема работы алгоритма трансформации поискового запроса в его унифицированное семантическое представление

3.3.2 Процесс поиска в многоязычных массивах, основанный на использовании метода трансформации поискового запроса

Далее рассмотрим алгоритм поиска документов в многоязычных массивах с использованием стандартных средств СУБД (Алгоритм 2):

Шаг 1. На вход поступает поисковый запрос, после чего он обрабатывается с помощью алгоритма 1.

Шаг 2. Средствами СУБД производится поиск наименований понятий запросов в многоязычном массиве (при поиске сравниваются не сами наименования понятий, а их номера в многоязычном словаре унифицированных формализованных представлений наименований понятий).

Шаг 3. Запускается процедура ранжирования результатов поиска, полученных с помощью стандартных средств СУБД. Процедура ранжирования зависит от типа поиска.

Шаг 4. Выдача результатов поиска пользователю.

На Рис. 3 представлена общая схема работы программного модуля, в котором реализованы описанные алгоритмы.

Целью эксперимента являлась проверка работоспособности предложенных методов поиска информации в многоязычном массиве, установление их эффективности [8], а также возможности их использования в промышленных информационных системах. Эксперимент проводился на основе разработанного авторами программного комплекса. В качестве исходных данных для эксперимента был взят массив текстов по тематике «Информационные технологии» (182641 текст).

Рисунок 3 Общая схема работы программного модуля поиска текстовой информации в многоязычных массивах

3.3.3 Эксперимент по проверке разработанного метода поиска в многоязычном массиве

Эксперимент проводился в несколько этапов:

1. На первом этапе тексты документов, приготовленные для эксперимента, были загружены в систему и обработаны при помощи *алгоритма 1*, изложенного в разделе 3.3.1. Все результаты обработки были занесены в базу данных программного комплекса.

2. На втором этапе из загруженных текстов было выбрано 35000 предложений и 90000 случайных наименований понятий. При этом был создан контрольный массив, где содержались все адреса предложений и наименований понятий в текстах документов коллекции.

3. На третьем этапе выбранные предложения и наименования понятий были переведены на английский и немецкий язык с помощью системы перевода Google Переводчик (https://translate.google.ru/).

4. На четвёртом этапе был произведен поиск каждого из переведенных на третьем этапе предложений и наименований понятий в русскоязычном массиве документов. Для этого использовался *алгоритм 2*, изложенный в разделе 3.3.2. После этого информация об адресах найденных соответствий сопоставлялась с информацией, полученной в п. 2.

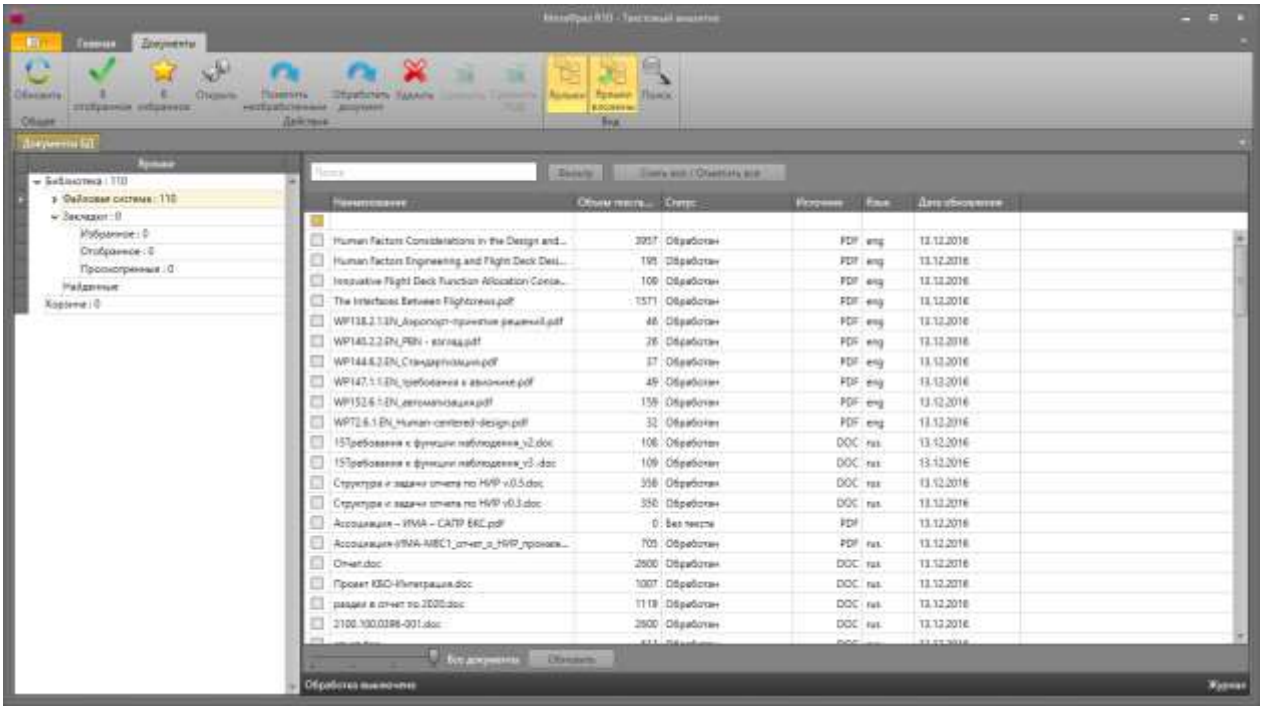


Рисунок 4 Скриншот интерфейса электронной библиотеки MF Text Analyst

5. На пятом этапе с помощью данных, полученных в п. 4, были получены значения полноты, точности и F1-меры. Результаты приведены в таблице 2.

Таблица 2 Значения показателей эффективности метода

	Полнота	**Точность**	**F1-мера**
Поиск наименований понятий	0.88	0.96	0.92
Поиск предложений	0.79	0.99	0.89
Среднее значение	0.84	0.98	0.91

4 Заключение

Идеи, описанные выше, были реализованы в виде программного продукта MF Text Analyst на базе программно-лингвистической платформы MetaFraz R10. Данный программный комплекс предназначен для выполнения следующих простых операций:

- ведение электронной библиотеки научно-технических документов;
- автоматическое формирование формализован ного представления документов;
- семантический поиск, отбор и сравнение документов.

MF Text Analyst позволяет загружать в БД документы в наиболее распространенных форматах (PDF, DOC, DOCX, TXT и др.), а затем извлекать текстовое содержимое и производить все этапы лингвистической обработки. Скриншот интерфейса электронной библиотеки MF Text Analyst представлен на рис. 4.

Также в данном программном продукте в тестовом режиме реализован многоязычный поиск. Его эффективность была проверена на коллекции размером в 182641 документ и показала неплохие для данного этапа исследований результаты. Предложенный авторами метод показал соответствующую аналогам скорость поиска при использовании СУБД RavenDB. Далее для улучшения показателей эффективности необходимо продолжать работу по доработке программного обеспечения, а также пополнять словари новой лексикой. Указанные мероприятия позволят значительно улучшить качество работы разработанных алгоритмов на текстах, относящихся к широкому спектру предметных областей.

Литература

[1] Елизаров, А.М., Зуев, Д.С, Липачёв, Е.К.: Свободно распространяемые системы управления электронными научными журналами и технологии электронных библиотек. Труды XV Всерос. науч. конф. «Электронные библиотеки: перспективные методы и технологии, электронные коллекции» – RCDL'2013, г. Ярославль, 14–17 октября 2013 года, сс. 227-236 (2013)

[2] Малахов, Д.А., Сидоренко, Ю. А., Атаева, О.М., Серебряков, В.А.: Семантический поиск как средство взаимодействия с электронной библиотекой. Труды XVIII Межд. конф. DAMDID / RCDL'2016 «Аналитика и управление данными в областях с интенсивным

использованием данных», 11–14 октября 2016 года, Ершово, Москва, сс. 85-91 (2016)

[3] Ле Хоай, Тузовский, А.Ф.: Разработка семантических электронных библиотек на основе онтологических моделей. Труды XV Всерос. науч. конф. «Электронные библиотеки: перспективные методы и технологии, электронные коллекции» – RCDL'2013, г. Ярославль, 14 – 17 октября 2013 года, сс. 143-151 (2013)

[4] Голицына, О.Л., Максимов, Н.В., Окропишина, О.В., Строгонов, В.И.: Онтологический подход к идентификации информации в задачах документального поиска: практическое применение. Научно-техническая информация. Серия 2: Информационные процессы и системы, (3), сс. 1-8 (2013)

[5] Добров, Б.В., Лукашевич, Н.В.: Организация двуязычного поиска в университетской информационной системе «Россия». Труды четвертой Всерос. науч. конф. RCDL'2002 «Электронные библиотеки: перспективные методы и технологии, электронные коллекции», г. Дубна, 15–17 октября 2002 г., сс. 148-158 (2002)

[6] Oard, D.: Alternative Approaches for Cross-Language Text Retrieval. Proc. of the AAAI Spring 1997 Symposium on Cross-Language Text and Speech Retrieval (1997)

[7] Cardeñosa, J., Gallardo, C., Toni, A.: Multilingual Cross Language Information Retrieval: A New Approach. Seventh Int. Conf. on Computer Science and Information Technologies, 28 September – 2 October, 2009, Yerevan, Armenia (2009)

[8] Хорошилов, А.А.: Методы автоматического установления смысловой близости документов на основе их концептуального анализа. Труды XV Всерос. науч. конф. «Электронные библиотеки: перспективные методы и технологии, электронные коллекции» – RCDL'2013, г. Ярославль, 14–17 октября 2013 года, сс. 369-376 (2013)

[9] Захаров, В.Н., Хорошилов, Ал-др А., Хорошилов, Ал-ей А.: Метод автоматического выявления неявно выраженных заимствований в научно-технических текстах. Искусственный интеллект и принятие решений, (1), сс. 10-20 (2017)

[10] Захаров, В.Н., Хорошилов, Ал-др. А., Хорошилов, Ал-ей. А.: Метод выявления заимствований в текстах разноязычных документов. Труды XVIII Межд. конф. DAMDID / RCDL'2016 «Аналитика и управление данными в областях с интенсивным использованием данных», 11–14 октября 2016 года, Ершово, Москва, сс. 277-282 (2016)

[11] Борзых, А.И., Брагина, Г.А., Хорошилов, А.А.: Методы автоматической кластеризации документов в хранилищах научно-технической информации для решения задачи поиска плагиата в текстах документов. Информатизация и связь, (8), сс. 33-37 (2012)

[12] Дмитришин, А.Н., Калинин, Ю.П., Никитин, Ю.В., Хорошилов, А.А., Хорошилов, А.А.: Технологии автоматической обработки и семантического анализа разноязычных документов в системе мониторинга мирового потока научно-технической информации крупного информационного центра. Информатизация и связь, (1), сс. 49-55 (2017)

[13] Zakharov, V., Khoroshilov, A.: Automatic Assessment of Similarity of the Texts' Thematic Content on The Base of their Formalized Semantic Descriptions Comparison. CEUR Workshop Proceedings. Proc. of the 14th All-Russian Scientific Conf. "Digital libraries: Advanced Methods and Technologies, Digital Collections", Pereslavl-Zalessky, Russia, October 15–18, 934, pp. 143-149 (2012)

[14] Zakharov, V., Khoroshilov, A.: Semantic Methods for Solving a Problem of Automatic Detection of Plagiarism in Structured Scientific and Technical Documents. CEUR Workshop Proceedings. Selected Papers of the 15th All-Russian Scientific Conf. "Digital Libraries: Advanced Methods and Technologies, Digital Collections", Yaroslavl, Russia, October 14–17, 1108, pp. 165-172 (2013)

[15] Khoroshilov, A.A.: Method for Detecting Implicit Plagiarism in Scientific and Technical Texts on the Basis of Their Conceptual Analysis. CEUR Workshop Proceedings. Selected Papers of the XVII Int. Conf. on Data Analytics and Management in Data Intensive Domains (DAMDID/RCDL 2015), Obninsk, Russia, October 13–16, 2015, 1536, pp. 266-372 (2015)

[16] Zakharov, V., Khoroshilov, Alexandr, Khoroshilov, Alexey: A Method of Automatic Plagiarism Detection in Multilingual Documents. CEUR Workshop Proceedings. Selected Papers of the XVIII Int. Conf. on Data Analytics and Management in Data Intensive Domains (DAMDID/RCDL 2016), 1752, pp. 181-186 (2016)

Digital Mathematical Libraries: Overview of Implementations and Content Management Services

© A.M. Elizarov © E.K. Lipachev © D.S. Zuev

Volga Region Federal University,
Kazan, Russia

amelizarov@gmail.com elipachev@gmail.com dzuev11@gmail.com

Abstract. The paper gives a review of existing projects of implementation of digital mathematical libraries. An analysis of existing information systems of digital mathematical libraries is performed using the evaluation criteria embedded in the DELOS DLRM model, emphasis is placed to the methods of managing mathematical content on the basis of semantic technologies. All projects are in different degrees of completeness, the range of services provided is different. We found that most of digital mathematical libraries are concentrated on the transfer of the resources to the electronic form and their preservation, rather than on the development of semantic services.

Keywords: digital publishing, library automation, machine-actionable digital library, digital mathematics library, DML, WDML.

1 Introduction

The Digital Era has changed crucially as the methods of research, and the ways in which scientists search, produce, publish, and disseminate their scientific work. A digital library, a collection of information which is both digitized and organized, gives us power we never had with traditional libraries. Information and communication technologies are actively implemented in research and development. Therefore, it became possible to use the entire volume of accumulated scientific knowledge in conducting new research. This requires creation of complex of technologies that ensure management of available knowledge, the organization has effective access to this knowledge, as well as sharing and multiple use of new kinds of knowledge structures. In mathematics also accumulated considerable experience in using of electronic mathematical content within the various projects on creation of mathematical digital libraries (see, e. g., [1]).

Since inception of the first scientific information systems, mathematicians have been involved in the full cycle of software product development, from idea to implementation. Well-known examples are an open source system TEX and commercial systems Wolfram Mathematica and Wolfram Alpha, led by Stephen Wolfram according to his principles of computational knowledge theory [2, 3]. Tools for mathematical content management are developed with the help of communities of mathematicians, e.g. MathJax by American Mathematical Society, information system Math-Net.Ru is developed at the Steklov Mathematical Institute of the Russian Academy of Sciences [4] and the collection of publicly available preprints arXiv.org (https://arxiv.org/).

Main challenges of mathematical knowledge management (MKM) are discussed in [5–9], the most urgent tasks are outlined. Such tasks are: modeling representations of mathematical knowledge; presentation formats; authoring languages and tools; creating repositories of formalized mathematics, and mathematical digital libraries; mathematical search and retrieval; implementing math assistants, tutoring and assessment systems; developing collaboration tools for mathematics; creating new tools for detecting re-purposing material, including plagiarism of others' work and self-plagiarism; creation of interactive documents; developing deduction systems. The solution of this task requires formalization of mathematical statements and proofs [9].

At present, research activities in the field of mathematics are associated with the use of modern information technology (cloud, semantic, etc.). These technologies are used in research of distributed scientific teams, preparation and dissemination of mathematical knowledge in an electronic form. At present, a new type of digital library is being formed, connected with the integration of mathematical knowledge into the scientific information space, see. [1,10,11]. This type of information system is called Digital Mathematical Library (DML), a number of global projects are implemented, such as European Digital Mathematical Library or World Digital Mathematical Library [12–14]. More details about goals, functions and current results are listed below, in Section 3.

Implementation of digital mathematical libraries involves the development of special tools and continuous improvement of their functionality. An example is the Open Journal Systems (OJS, https://pkp.sfu.ca/ojs/). The platform is used in many projects, particularly in Lobachevskii Journal of Mathematics (http://ljm.kpfu.ru/), one of the first digital mathematical journals [15].

In our work, we try to look more deeply into world largest DML to outline current status of described projects and to investigate services and functions that provide these digital mathematical libraries.

Proceedings of the XIX International Conference "Data Analytics and Management in Data Intensive Domains" (DAMDID/RCDL'2017), Moscow, Russia, October 10–13, 2017

2 Mathematical Libraries and DELOS Digital Library Reference Model

2.1 Criteria for investigation

Firstly, we need to establish common criteria and main features and functions that we will look at.

In DELOS Digital Library Reference Model [16, 17] three basic concepts are distinguished for defining what is called a digital library (DL):

• DL – a (potentially virtual) organization that comprehensively collects, manages, and preserves for the long term rich digital content and offers to its user communities specialized functionality on that content, of measurable quality, and according to prescribed policies;

• DL system – a software system that is based on an architecture and provides all functionality that is required by a particular Digital Library. Users interact with a Digital Library through the corresponding DL system;

• DL management system (DLMS) – a generic software system that provides the appropriate software infrastructure to both produce a basic DL system that incorporates all functionality that is considered foundational for Digital Libraries and integrate additional software offering more refined, specialized, or advanced functionality. An intrinsic part of DLMS functionality is related to administrative services that are used to choose the appropriate subset of its functionality, e.g., through relevant parameters of its components, and then install, deploy, and (re)configure a DL system.

A DLMSis "system software". As in several other domains (e.g., operating systems, databases, user interfaces), such kernel software may be used as a foundation to produce Digital Library systems.

While the concept of DL is intended to capture an abstract system that consists of both physical and virtual components, the remaining two capture concrete software systems. For every DL, there is a unique DL system in operation (possibly consisting of many interconnected smaller DL systems in the most general case), where as all DL systems are based on a handful of DLMSs.

In the role-based aspect, the DELOS DLRM model consider following types of users: the end user of the DL; the developer of DL; the system administrator of DL and the developer of applications for DL and, four levels of user views and expectations are formed. In addition, the model identifies six key areas, each of which introduces and defines its own entities and their properties: architecture, information space, functionality, users, policy and quality of services provided. These areas can be considered as evaluation criteria and, by virtue of their universality, can be used to analyze almost any information system.

We will carry out an analysis of existing digital mathematical libraries, performed using the evaluation criteria embedded in the DELOS DLRM model.

2.2 Differences between approaches

It is interesting to stop at the discussion at the approaches of the definition of elementary objects with which digital library works. In particular, an interesting approach to information objects organization lies in the ideology of WDML. We use the same approach in creating a digital mathematical library Lobachevskii DML, which is based on mathematical collections of the Kazan Federal University [18].

Usually digital library consists of collections, and collections in turn from documents or information resources (objects). In 1990–2000 there was a large number of studies carried out on the definition, architectural and technical aspects of DL systems. Finally, it is necessary to mention the creation of the DL manifesto in the DELOS project, which resulted in the creation of a reference model for DL [16, 17].With the development of Semantic Web technologies, it became interesting to investigate the semantics of resources and their links placed in libraries, see, for example, [18]. In this case, an information object can already be considered not only as a document, but as its certain parts – abstract, keywords, bibliography, citations, comments of authors or readers.

From the end user's point of view, DL must satisfy the user's expectations. The document itself as an elementary information object may not be interesting at all. It is much in demand to search for information on a particular entity or subject mentioned in the document. At the same time, much more interesting to find all possible resources where different versions of mentioned subjects, especially in cases when various interpretations and definitions are possible. For example, there are a number of definitions of the concept of "digital library" and the user studying this topic will certainly be interested in all references to the definition of the digital library from different sources. Thus, we observe a change in the elementary information object. The electronic document fragmented into smaller information objects and all services of a library deal with such objects and manage the relationships between them. In mathematics, such elementary objects can be, for example, theorems, lemmas, definitions or formulas, research of which is much more informative on a number of sources. The services of any DML should provide such an opportunity. All this functionality lies in WDML architecture. Its implementation became possible only with the development of semantic technologies and the transfer of library content into digital form with metadata. Now, there is no technical problems in maintaining such approach to the organization of DL.

During our research, we will take into account this transformation of the approach to the organization of information objects. Note that the change in the approach to the organization of DL does not affect the selected criteria for investigation.

3 Functionality of Digital Mathematical Libraries

Below is a brief review of existing digital mathematical libraries. The largest projects are "All-Russian Mathematical Portal Math-Net.RU","Centre de diffusion de revues académiques mathématiques", "Czech Digital Mathematics Library", "The Polish Digital Mathematics Library", "Göttinger

Digitalisierungs Zentrum", "Numérisation de documents anciens mathématiques", Zentralblatt MATH, "Bulgarian Digital Mathematics Library" and "The European Digital Mathematical Library". It should be outlined, that all projects are in different degrees of completeness, the range of services provided is also different.

3.1 Math-Net.ru

All-Russian Mathematical Portal Math-Net.ru [4, 20–22] combines both a digital mathematical library and a publishing system for mathematical texts. It is a web portal developed by the V. A. Steklov Institute of Mathematics, Russian Academy of Sciences.

The key component of the portal – the "Journals" section links Russian periodicals in the field of mathematical sciences to a single information system. Currently contains more than 120 journals with nearly 200 thousand publications. Information about the article includes a bibliographic description, an annotation, lists of literature and a file with the full text of the article. The portal presented in two languages – Russian and English.

The most interesting part is the functionality of the portal. The portal provides the ability to search for publications and links on the bibliographic description and keywords in the title, annotation or text. As result of the search, an abstract, article IDs (DOI, resource references in abstract databases, URIs), a citation pattern, classifier values are issued. There are no recommender service, in fact all semantic services work with a bibliographic description of the resource. MiRef module is used to form correctly the description and links to resources. The module is designed to automatically place links to various publications databases in the literature list. The format of the links must satisfy the rules of the amsbib package and should be entered in the LaTeX format.

Registered users can create personal pages, manage personal collections of publications, authors get access to the full texts of their articles, authors can send the manuscript to the editorial office of the journal electronically, and track the process of its workflow in the editorial office.

Statistics on popular authors and resources are maintained, infometric indicators for resources located on the portal are calculated.

The policy for accessing the full texts of articles is determined by the publisher of the paper. Access for any other information is free.

3.2 CEDRAM

The center for diffusion of academic mathematical journals (CEDRAM, Centre de Diffusion de Revues Académiques Mathématiques) is a web portal for common access to a set of mathematical journals [23], available in French and English. CEDRAM's mission is to provide a large distribution of their current volumes, and range from help for producing journals according to the best standards for electronic publishing to long lasting archiving. CEDRAM is a service of the Cellule MathDoc (UMS 5638 of CNRS and Université Joseph Fourier) which completes its important offer in mathematical documentation. This DML is not so large – contains 9 French math journals, 1 book and 7 proceedings of seminars and conferences.

The CEDRAM websites offer two ways of consulting the hosted articles: quick and advanced search. Search functions provide search by keywords, author, title, bibliography and full text search. Quick search searches in all fields except full text. Advanced search interface offers several types of research, more or less complicated. The full entry of articles produced for CEDRAM contains abstracts and bibliographical references.

All online records exist in two formats, which are only different by the way they display mathematical formulas in titles, abstracts, keywords or references: MathML or TeX and have stable url link.

XHTML+MathML display is best for reading and browsing, but there are some problems with viewing in browsers, that need to be pre-configured to work correctly with MathML. The HTML+TeX version used for compatibility for users who do not have an environment capable of displaying MathML. Now CEDRAM provide following services [12, 13, 23]:

- production workflow of journals;
- dedicated web site for each journal;
- provides creation and maintenance of LATEX styles (using a specific class);
- production of PDFs for print and web with XML/MathML metadata;
- DOI registration (Crossref), reference linking (MSN, ZBM, mini-DML, Crossref);
- provides publishing platform for mathematical articles based on Open Journal System (Public Knowledge Project, https://pkp.sfu.ca/ojs/);
- all resources archived in partner project - the French digital math library NUMDAM.

Policy and quality of services. Starting 2017 all CEDRAM journals are open access. Access to the database containing the bibliographical references of all the articles of all participating journals is totally free. The database itself is the property of Cellule Mathdoc, and contains elements covered by copyright. CEDRAM has OAI-PMH server, which can be used for systematic download of metadata in various schemas. Files of the full texts are the property of the journals and it is necessary to refer to the policy of each of them. Also there are some restrictions of full copying and indexing by web robots.

3.3 Numerisation de Documents Anciens Mathematiques (NUMDAM)

The French digital math library NUMDAM [12–14, 24] started as a digitisation program for a pilot of 6 journals. Now it contains more than 57000 articles in 76 periodicals, 373 books in 4 collections, 263 theses.

The NUMDAM is the reference French digital mathematics library set up by Cellule MathDoc with the assistance of a network of partners.

From 2007 onwards, publishers send digital born articles into DML. Collections are normally indexed

within one year of publication, and full texts are freely downloadable at the end of a period of time set by agreement upon each title.

The NUMDAM program is designed to support academic publishers and provide the research community with a sustainable, reliable and easy-to-use library. The research and dissemination platform was completely redesigned in 2016. Now portal is available on two languages –English and French, formulas can displayed in TeX or in graphical form using MathJax.

System provide following functions: search and navigation by title, author, references or in full text of resources. During search all statistics, related to the search topic is displayed – co-authors, journals and years of publication. Browse functions provide navigation through sorted list of resources (authors, journals etc.).

Full texts available in PDF and DJVU formats. Each article in NUMDAM is available via a stable URL. This URL is a compact address, designed to remain valid in the long term. It is displayed in the web page of the article, on the first page of PDF or DjVu files and by the OAI-PMH server.

There is no any user registration. All functions have open access. NUMDAM only disseminate resources that already published in journals, books or theses but submission process of resources is not clear. Metadata extraction made only for bibliography. Any additional services like formula search or recommender system are absent.

The full text of most recent articles is generally not available. The journals whose archives are on this portal have accepted the principle of "a moving wall". This is a time interval between the publication of a volume (in paper or electronic form, delivered to subscribers) and the availability of the full text on the NUMDAM server. Generally, moving-wall for most of journal in NUMDAM is equal to 5 years.

3.4 The Czech Digital Mathematics Library (DML-CZ)

The Czech Digital Mathematics Library (DML-CZ) [25, 26] has been developed in order to preserve in a digital form the content of major part of mathematical literature that has ever been published in the Czech lands, and to provide a free access to the digital content and bibliographical data. DML-CZ resulted from the project no. 1ET200190513 supported by the Czech Academy of Sciences (CAS) in the R&D programme Information Society, and operated by the Institute of Mathematics CAS. Project seems to be finished in 2010 and now is in stable form.

Functionality. Editors of all journals included in DML-CZ are using tools and work flows that have been tailored to their individual publishing practice and that enable them to produce inputs for DML-CZ in a semiautomatic way. The formal consistency and integrity of the data are controlled by several validating procedures that have been developed in the project.

There are some automated procedures for validation of data of new journal issues but all of them are archived in DML-CZ for internal use and development. Based on limiting the name space of allowed TEX macros, validation service get all metadata including abstracts, keywords and references transformed into representation using MathML [27].

End-users cannot submit any resource, everything can be submitted only through editorial board of journals, also there is no any personal area for users.

Search and navigation. As others DMLs DML-CZ allows to search by title, author of publications. Also avaliable search by language or by Zentrablatt MATH and MathSciNet identifiers. Browse functions provide navigation through sorted list of resources (authors, journals etc.).

The most interesting function is search of related articles (finding similarities between papers). This service tries to find similar papers using three methods: "Term frequency–Inverse document frequency" (TF-IDF, see, e. g. [28]), the "Random Projections" or method that is built on TF-IDF and simplifies the computations by projecting vectors onto a subspace of lower dimensionality [28] and with using "Latent Semantic Indexing" (LSI, [29]). Last method gives the most accurate results up to 90%.

Policies and quality of service. The digitized journal and proceedings papers are displayed with the agreement of the publisher who owns the digital data. The digitized monographs are displayed with the agreement of the author and/or the publisher while the digital data are property of the Institute of Mathematics CAS. The database itself, in particular the bibliographic data, are property of the Institute of Mathematics CAS.DML-CZ presents full texts articles and book chapters in PDF format, equipped with enhanced metadata including bibliographical references linked to Zentrablatt MATH and MathSciNet. The digital born documents are being obtained from the original sources provided by publishers. The presented page content and format corresponds to the original one. Journals are presented and accessed according to the terms of a contract with the publisher. The digital documents displayed in the DML-CZ are authorized with electronic stamps.

3.5 The Polish Digital Mathematical Library

The Polish Digital Mathematical Library (DML-PL, [30]) has existed since 2002. The library holds full texts of polish mathematical journals and books. The major part of the collection are archive issues of mathematical journals published before World War II. Library consists of 550 books and 36 journals, but only 3 journals provide access to full text of articles. Portal of DML-PL provide search by attributes and navigation through sorted lists of authors, books and journals.

Brief explanation of the project is given in [31], but nowadays it seems that project is already finished. On the web portal of library there is no additional information about current status. Any information about semantic functions or metadata extraction from resources is missing.

3.6 GDZ–Gottingen Digitization Centre

The task of the GDZ [32, 33] is to record data such as prints, manuscripts and illustrations and to preserve them. Main aim of the project is conversion of resources

into digital form. This is multidisciplinary library, that contains not only mathematical collections but also history of Law, history of the Humanities and the Sciences, travel and North American literature and other collections. Mathematical collections have about 7000 resources and also have some Russian resources. Library contains more than 15 million digitized pages.

Portal provides search in metadata and full text of resources and browse functions. Many resources are historical, not modern, main aim of the project is to digitize and preserve resources. All resources have full texts and can be viewed page by page or in structured mode. Metadata of any resource contain stable URL of resource, metadata can be downloaded in METS format.

3.7 Zentralblatt MATH

Zentralblatt MATH (zbMATH, [34]) is abstracting and reviewing service in pure and applied mathematics. It is hosted by the Berlin office ofFIZ Karlsruhe – Leibniz Institute for Information Infrastructure GmbH (FIZ Karlsruhe) and distributed by Springer. The zbMATH database contains more than 3.5 million bibliographic entries with reviews or abstracts currently drawn from more than 3,000 journals and serials, and 170000 books. zbMATH is not a digital library itself, it is an indexing service and provides easy access to bibliographic data, reviews and abstracts from all areas of pure mathematics as well as applications, in particular to the natural sciences, computer science, economics and engineering.

Search functions provide search for documents, authors and journals. Search can be done in one line, or in structured form using attributes such as title, author, subject, source, keywords etc. Service also provide full-text formula search for indexed arXiv documents [35].The zbMATH formula search uses the MathWebSearchsystem, which is a content-based search engine for MathML formula based on substitution tree indexing.

Portal offer three ways of displaying mathematical formulas – MathML, MathJax and LaTeX. The XML-based MathML is the solution recommended by W3C for displaying mathematical content on the web and is set as default within zbMATH. Mathematical Reviews and zbMATH maintain the Mathematics Subject Classification (MSC), a classification scheme for mathematics. It is used by reviewing services to categorize items in the mathematical sciences literature. The database of service contains about 2.1 million direct links to electronic versions of the indexed publications, to the publishers' websites and/or to electronic libraries with open access to the full texts.

3.8 Bulgarian Digital Mathematics Library

Bulgarian Digital Mathematics Library, BulDML is a digital repository at Institute of Mathematics and Informatics of Bulgarian Academy of Sciences. Library has 7 mathematical journals, 4 book series and proceedings in its repository. In fact, BulDML is an institutional repository and is built on open-source DSpace software [36]. As known, DSpace preserves and enables open access to all types of digital content including text, images, moving images, mpegs and data sets. All functionality of DSpace software is clear and we will not describe it in this paper. For example, additional information about DSpace can be found in [17, 37].

3.9 European Digital Mathematics Library

The European Digital Library (EuDML) was a project partly funded by the European Commission. EuDML [12–14, 38, 39] is an aggregation and indexing services with was established under The EuDML Initiative and promoted by European Mathematical Society. EuDML assemble as much as possible of the digital mathematical corpus in order to make it available online, with eventual open access, in the form of an authoritative and enduring digital collection, growing continuously with publisher supplied new content, augmented with sophisticated search interfaces and interoperability services, developed and curated by a network of institutions.

The system, presented in the diagram in Figure 1, conceptually consists of a metadata repository, a search engine, a metadata enhancer, an association analyser, annotation and accessibility functions and of course the interfaces [38].

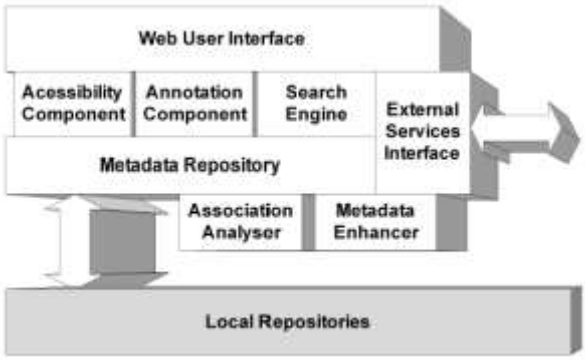


Figure 1 EuDML architecture

The metadata repository provides the central point of reference for all the managed contents. It will work with an OAI-PMH harvester to ingest repositories' content descriptions, maps the metadata into the internal EuDML schema. The performance and the quality of responses of the search service directly influence user experience. Therefore, particularly this service has to be reliable, scalable and customized to fulfill user expectations.

The metadata enhancer function consist in a collection of tools that each contribute to expand or complete the existing items' metadata, depending on the improvements needed. These range from applying OCR over full texts, adding key words or multilingual metadata by merging information from different databases when an item happens to have such non-redundant description, generating MathML for mathematical expressions, etc. The association analyzer detects, analyses and records relations between individual items. The annotation component provides mechanisms to attach new material to individual items in the repositories and maintain this new material. The accessibility component provides support for enhanced accessibility of items, if required, before presentation to end users. Finally, the user and system interfaces provide

access to the collected resources on different levels both to human and machine users. Now EuDML offers several service interfaces that allow other applications to connect with the service. These are OAI-PMH server, REST services, OpenSearch service, which allow to query library index in machine way and annotation retrieval services in JSON.

EuDML aims to be an open source of trusted mathematical knowledge. That is why it has some policies:

- All texts must have been scientifically validated and formally published;
- All items must be open access after a finite embargo period. Once documents contributed to the library are made open access due to this policy, they cannot revert to close access later on;
- The digital full text of each item contributed to library must be archived physically at one of the EuDML member institutions.

All DMLs, described above except All-Russian Mathematical Portal Math-Net.RU are partners of EuDML.

4 Conclusion

In order to outline all differences of observed projects we created comparison Table 1 listed below. Note that, we excluded from table two DMLs due to following. BulDML is and built on open-source DSpace software, so all functionality of it is clear, for DML-PL we could not find any working portal in order to study it more deeply.

In all the projects studied, emphasis is placed on the transfer of the resources themselves to the electronic form, rather than on the development of semantic services. Only a few portals have a mathematical formula search, and only one has a recommender service.

After the analysis done it is clear that there are only two types of repository systems: the first is actually DML, which preserve the resources themselves, the second is indexing and aggregating services that do not have their own database of electronic documents, but provide a wide range of convenient search capabilities.

This work was funded by the subsidy allocated to Kazan Federal University for the state assignment in the sphere of scientific activities, grant agreement no. 1.2368.2017) and with partial financial support of the Russian Foundation for Basic Research and the Government of the Republic of Tatarstan, within the framework of scientific projects Nos. 15-07-08522, 15-47-02472.

References

[1] Borwein, J.M., Rocha, E.M., Rodrigues, J.F. Communicating Mathematics in the Digital Era, pp. 3-21. A K Peters, Ltd. MKM-IG. Mathematical Knowledge Management (2008). http://www.mkm-ig.org/

[2] Wolfram, S.: A New Kind of Science. Wolfram Media, Inc. (2002)

[3] Wolfram, S.: An elementary introduction to the Wolfram Language. Wolfram Media, Inc. (2015)

[4] Chebukov, D.E., Izaak, A.D., Misyurina, O.G., Pupyrev, Yu.A., Zhizhchenko, A.B.: Math-Net.Ru as a Digital Archive of the Russian Mathematical Knowledge from the XIX Century to Today. Intelligent Computer Mathematics, Lecture Notes in Comput. Sci., 7961, pp. 344-348, Springer (2013), doi: 10.1007/978-3-642-39320-4_26

[5] Carette, J., Farmer, W.M.: A Review of Mathematical Knowledge Management. In Intelligent Computer Mathematics. Lecture Notes in Computer Science, 5625. pp. 233-246 (2009)

[6] Ion, P.D.F.: Mathematics and the World Wide Web. In Intelligent Computer Mathematics. Lecture Notes in Computer Science, 7961, pp. 230-245 (2013)

[7] Lange, C.: Enabling Collaboration on Semiformal Mathematical Knowledge by Semantic Web Integration. Ph. D. Thesis, Jacobs University Bremen (2011)

[8] Elizarov, A.M., Lipachev, E.K., Nevzorova, O.A., Solov'ev, V.D.: Methods and Means for Semantic Structuring of Electronic Mathematical Documents. Doklady Mathematics, 90 (1), pp. 521-524 (2014), doi: 10.1134/S1064562414050275

[9] Elizarov, A., Kirillovich, A., Lipachev, E., Nevzorova, O., Solovyev, V., and Zhiltsov N.: Mathematical Knowledge Representation: Semantic Models and Formalisms. Lobachevskii J. of Mathematics, 35 (4), pp. 347-353 (2014), doi:10.1134/S1995080214040143

[10] Elizarov A., Kirillovich A., Lipachev E., Nevzorova O. (2017) Digital Ecosystem OntoMath: Mathematical Knowledge Analytics and Management. In: Kalinichenko L., Kuznetsov S., Manolopoulos Y. (eds) Data Analytics and Management in Data Intensive Domains. DAMDID/RCDL 2016. Communications in Computer and Information Science, 706, pp. 33-46 (2017), doi: 10.1007/978-3-319-57135-5_3

[11] Elizarov, A.M., Kirilovich, A.V., Lipachev, E.K., Nevzorova, O.A.: Mathematical Knowledge Management: Ontological Models and Digital Technology. CEUR Workshop Proceedings, 1752, pp. 44-50 (2016), http://ceur-ws.org/Vol-1752/paper08.pdf

[12] Bouche, T.: Towards a World Digital Library: Mathdoc, Numdam and EuDML Experiences. UMI, La Sapienza, Roma (2016), http://www.mat.uniroma1.it/sites/default/import-files/biblioteca/SEMINARIO2016/bouche.pdf

[13] Bouche, T.: Digital Mathematics Libraries: The good, the bad, the ugly. Mathematics in Computer Science, (3), pp. 227-241 (2010), doi: 10.1007/s11786-010-0029-2

[14] Bouche, T.: Reviving the Free Public Scientific Library in the Digital Age? The EuDML Project. In: Kaiser, K., Krantz, S., Wegner, B. (eds.): Topics and Issues in Electronic Publishing, JMM, Special

Session, San Diego, January 2013, pp. 57-80 (2013), http://www.emis.de/ proceedings/TIEP2013/05bouche.pdf

[15] Elizarov, A.M., Zuev, D.S., Lipachev, E.K.: Mathematical Content Semantic Markup Methods and Open Scientific E-Journals Management Systems. In: Klinov, P., Mouromtsev, D. (eds.) KESW 2014. CCIS, 468, pp. 242-251 (2014), doi: 10.1007/978-3-319-11716-4 22 29

[16] Candela, L., Athanasopoulos, G., Castelli, D., El Raheb, K., Innocenti, P., Ioannidis, Y., Katifori, A., Nika, A., Vullo, G., Ross, S.: The Digital Library Reference Model. FP7-ICT-2007-3. Cultural Heritage and Technology Enhanced Learning (2011)

[17] Candela, L., Castelli, D., Fuhr, N., Ioannidis, Y., Klas, C.-P., Pagano, P., Ross, S., Saidis, C., Schek, H.-J., Schuldt, H., Springmann, M.: Current Digital Library Systems: User Requirements vs Provided Functionality. IST-2002-2.3.1.12. Technology-enhanced Learning and Access to Cultural Heritage (2006)

[18] Elizarov, A.M., Lipachev, E.K.: Lobachevskii DML: Towards a Semantic Digital Mathema-tical Library of Kazan University, 2017 (in press), DAMDID-2017 proceedings

[19] Kogalovskiy, M.R., Parinov, S.I.: Klassifikatsiya i ispol'zovaniye semanticheskikh svyazey mezh-du informatsionnymi ob'yektami v nauchnykh elektronnykh bibliotekakh. Inform. i yee primen., 3 (6), pp. 32-42 (2012)

[20] All-Russian Mathematical Portal Math-Net.Ru. http://www.mathnet.ru/

[21] Zhizhchenko, A.B., Izaak, A.D.: The Information System Math-Net.Ru. Application of Contemporary Technologies in the Scientific Work of Mathematicians. Russian Math. Surveys, 62 (5), pp. 943-966 (2007), http://dx. doi.org/10.1070/ RM2007v062n05ABEH004455

[22] Zhizhchenko, A.B., Izaak, A.D.: The Information System Math-Net.Ru. Current State and Prospects. The Impact Factors of Russian Mathematics Journals. Russian Math. Surveys, 64 (4), pp. 775-784 (2009), http://dx.doi.org/ 10.1070/ RM2009v064n04ABEH004638

[23] CEDRAM. www.cedram.org

[24] NUMDAM. www.numdam.org

[25] The Czech Digital Mathematics Library (DML-CZ), http://www.dml.cz/

[26] The Czech Digital Mathematics Library. Project Funded by the Academy of Sciences of the Czech Republic, 2005–2009. http://project.dml. cz

[27] Rákosník, J.: Recent Development of the DML-CZ and Its Current State. In Proc. of DML 2011: Towards a Digital Mathematics Library. Bertinoro, Italy, July 20–21st (2011)

[28] Rajaraman, A.; Ullman, J. D.: Data Mining (2011). doi:10.1017/CBO9781139058452.002

[29] Deerwester, S., Dumais, S., Landauer, T., Furnas, G., Beck, L.: Improving Information Retrieval with Latent Semantic Indexing. Proc. of the 51st Annual Meeting of the American Society for Information Science, 25, pp. 36-40 (1988)

[30] The Polish Digital Mathematics Library, http://pldml.icm.edu.pl/

[31] Zamlynska, K., Tarkowski, A., Rosiek, T.: Evolution of the Mathematical Collection of the Polish Virtual Library of Science. Mathematics in computer Science, (3), pp. 265-278 (2010), doi: 10.1007/s11786-010-0029-2

[32] Gottingen Digitalisierungs Zentrum. http://gdz. sub.uni-goettingen.de/gdz/

[33] Gottingen digitization Centre https://www.sub. uni-goettingen.de/en/copying-digitising/ goettingen-digitisation-centre/

[34] Zentralblatt MATH. https://zbmath.org/

[35] Muller, F., Teschke, O.: Full Text Formula Search in zbMATH, EMS Newsletter (2016)

[36] Bulgarian Digital Mathematics Library. http://sci-gems.math.bas.bg/jspui/

[37] DSpace, www.dspace.org

[38] Sylwestrzak, W., Borbinha, J., Bouche, T., Nowinski, A., Sojka P.: EuDML – Towards the European Digital Mathematics Library. In: Sojka, P. (ed.) Towards a Digital Mathematics Library. Paris, July 7–8th, 2010, pp. 11-26. Masaryk University Press, Brno (2010), http://dml.cz/bitstream/handle/10338.dmlcz/702569/DML_003-2010-1_5.pdf

[39] EuDML, www.eudml.org

Table 1 Comparison table of DML projects

DML Criteria	Math-Net.ru	CEDRAM	NUMDAM	DML-CZ	GDZ	zbMATH	EuDML
Information space	There is an object hierarchy. Collections split into journals, issues, articles and so on. Currently contains more than 120 journals with nearly 200 thousand publications. Information about the article includes a bibliographic description, an annotation, lists of literature and a file with the full text of the article.	DML contains 9 French math journals, 1 book and 7 proceedings of seminars and conferences. All CEDRAM journals are open access. Access to the database containing the bibliographical references of all the articles of all participating journals is totally free. The full entry of articles contains abstracts and bibliographical references.	Contains more than 57000 articles in 76 periodicals, 373 books in 4 collections, 263 theses. Full texts available in PDF and DJVU formats. Each article in NUMDAM is available via a stable URL.	The digitized journal and proceedings papers are displayed with the agreement of the publisher who owns the digital data. DML-CZ presents full texts articles and book chapters in PDF format, equipped with enhanced metadata including bibliographical references. The digital born documents are being obtained from the original sources provided by publishers.	This is multidisciplinary library, that contains not only mathematical collections but also history of Law, history of the Humanities and the Sciences, travel and North American literature and other collections. Mathematical collections have about 7000 resources and also have some Russian resources. Library contains more than 15 million digitized pages.	The database contains more than 3.5 million bibliographic entries with reviews or abstracts currently drawn from more than 3,000 journals and serials, and 170,000 books. The database of service contains about 2.1 million direct links to electronic versions of the indexed publications, to the publishers' websites and/or to electronic libraries with open access to the full texts.	This is an aggregation and indexing service. EuDML assemble the digital mathematical corpus in order to make it available online.
Functionality	The portal provides the ability to search for publications and links on the bibliographic description and keywords in the title, annotation or text. As result of the search, an abstract, article IDs (DOI, resource references in abstract databases, URIs), a citation pattern, classifier values are issued. There are no recommender service, in fact all semantic services work with a bibliographic description of the resource	CEDRAM has OAI-PMH server, which can be used for systematic download of metadata in various schemas. Search functions provide search by keywords, author, title, bibliography and full text search. Quick search searches in all fields except full text. Advanced search interface offers several types of research, more or less complicated. The full entry of articles produced for CEDRAM contains abstracts and bibliographical references.	NUMDAM has an OAI-PMH server, thus allowing sharing of metadata and better visibility of collections. System provide following functions: search and navigation by title, author, references or in full text of resources. During search all statistics, related to the search topic is displayed – co-authors, journals and years of publication. Browse functions provide navigation through sorted list of resources. Metadata extraction made only for bibliography. Any additional services like formula search or recommender system are absent.	Editors of all journals are using tools and workflows that enable them to produce inputs in a semiautomatic way. The formal consistency and integrity of the data are controlled by several validating procedures that have been developed in the project. There are some automated procedures for validation of data of new journal issues but all of them are for internal use and development. DML-CZ allows to search by title, author of publications, by language or by zbMATH and MathSciNet identifiers. Browse functions provide navigation through sorted list of resources. There is search of related articles.	Portal provides search in metadata and full text of resources and browse functions. All resources have full texts and can be viewed page by page or in structured mode. Metadata of any resource contain stable URL of resource, metadata can be downloaded in METS format.	Search functions provide search for documents, authors and journals. Search can be done in one line, or in structured form using attributes. Service also provide full-text formula search for indexed arXiv documents. The zbMATH formula search uses the MathWebSearch system. zbMATH maintain a classification scheme for mathematics.	EuDML offers several service interfaces that allow other applications to connect with the service. These are OAI-PMH server, REST services, OpenSearch service, which allow to query library index in machine way and annotation retrieval services in JSON.

Users	There are role model of users, everybody can register and create own personal area. Registered users can create personal pages, manage personal collections of publications, authors get access to the full texts of their articles.	No any user registration	No any user registration	No any user registration. End-users cannot submit any resource, everything can be submitted only through editorial board of journals, also there is no any personal area for users.	No any user registration.	There is a personal area for users – for reviewers, publishers etc.	No any user registration.
Quality of service	System is available in two languages. The policy for accessing the full texts of articles is determined by the publisher of the paper. Access for any other information is free.	Portal is available in English and French. Files of the full texts are the property of the journals. All online records exist in two formats, which are only different by the way they display mathematical formulas: MathML or TeX and have stable url link.	Portal available on English and French. Formulas can be viewed in TeX source code or in compiled, graphical way. NUMDAM only disseminate resources that were already published in journals, books or theses but submission process of resources is not clear.	Project was finished in 2010 and now it is in a stable form. Portal is available only in English.	Portal is available in German and English. But main aim of the project is to digitize and preserve resources.	Portal offers three ways of displaying mathematical formulas – MathML, MathJax and LaTeX. MathML is set as default. Not all services of the system are free, some of them need to be purchased.	EuDML has some policies: all texts must be scientifically validated and formally published; all items must be open access after a finite embargo period. Once documents contributed to the library are made open access due to this policy, they cannot revert to close access later on; the digital full text of each item contributed to library must be archived physically at one of the member institutions.

Lobachevskii DML: Towards a Semantic Digital Mathematical Library of Kazan University

© Alexander M. Elizarov © Evgeny K. Lipachev

N. I. Lobachevskii Institute of Mathematics and Mechanics,
Kazan (Volga Region) Federal University,
Kazan, Russia

amelizarov@gmail.com elipachev@gmail.com

Abstract. The digital mathematical library Lobachevskii DML is one of the national initiatives that have emerged in the past decade in different countries of the world. During this time, the formed technical and organizational conditions allowed making mathematicians' dreams of a global World Digital Mathematical Library (WDML) a reality. Following the vision approved by the International Mathematical Union, we started the Lobachevskii DML project in 2017 – the year of the 225-th anniversary of the birth of the brilliant mathematician Nikolai Ivanovich Lobachevskii, the founder of non-Euclidean geometry, the rector of the Kazan University. The main task of Lobachevskii DML project is the development of tools for managing mathematical content, which take into account not only the specifics of mathematical texts, but also the features of processing Russian-language texts. A particular task of creating this digital library is the integration of mathematical resources of Kazan University. Therefore, the original goal of the project was to build up a sound basis for a digital archive comprising the relevant mathematical literature published for 213 years of the existence of Kazan University and stored in the libraries of the University and Kazan. According to our assumption, the digital library Lobachevskii DML should be endowed with all conceivable necessary functions and services, making it a comprehensive and up-to-date live DML, generally respected and used by the local as well as the global mathematical community. From the very beginning, we had in mind that the Lobachevskii DML should constitute a building block for the envisioned global WDML.

In this paper, the results of the implementation of the digital mathematical library Lobachevsky-DML are presented. We describe the purpose of creating this digital library, methods of managing mathematical content based on semantic technologies. The following show how Lobachevskii DML interacts with the information systems of scientific journals. We also present a system of services to support the life cycle of a mathematical document and highlight technologies for supporting new forms of scientific publications and providing integration services with other digital mathematical archives and libraries.

Keywords: semantic technologies, semantic publishing, digital mathematics library, DML, World Digital Mathematics Library (WDML) project, Lobachevskii DML

Introduction

The creation and development of specialized digital libraries (DL) are one of the directions for the formation of a global scientific infrastructure. In the field of mathematics, the problems of integrating knowledge obtained over the entire "printed" period of the development of this science have been considered in a number of projects (see, for example, [1]). Even when such projects were of a local nature, the methods and tools developed during their implementation were oriented towards a comprehensive integration of knowledge (see, for example, [2–4]).

The modern vision of the tasks of forming a global infrastructure of mathematical knowledge is reflected in the documents of the World Digital Mathematics Library (WDML) project [5]. In these documents, it was noted that the leading role in the formation of digital mathematical collections, the accompaniment of their metadata (annotations, key words, etc.) is given by the "smaller" DML.

The present work is devoted to the presentation of Lobachevskii Digital Mathematics Library (Lobachevskii DML, http://www.Lobachevskii-dml.ru/), which we develop in accordance with the basic principles of WDML. The digital mathematical library Lobachevskii DML is another of the national initiatives that have emerged in the past decade in different countries of the world. During this time, the formed technical and organizational conditions allowed making mathematicians' dreams of a global WDML a reality. Following the vision approved by the International Mathematical Union, we started the Lobachevskii DML project in 2017 – the year of the 225-th anniversary of the birth of the brilliant mathematician Nikolai Ivanovich Lobachevskii, the founder of non-Euclidean geometry, the rector (from 1827 to 1845) of the Kazan University. The year 2017 was announced at the Kazan University of as the "Year of N. I. Lobachevskii".

The main task of Lobachevskii DML project is the development of such tools for managing mathematical content, which take into account not only the specifics of mathematical texts, but also the features of processing Russian-language texts. A particular task of creating this digital library is the integration of mathematical resources of Kazan University. Therefore, the original

goal of the project was to build up a sound basis for a digital archive comprising the relevant mathematical literature published for 213 years of the existence of Kazan University and stored in the libraries of the University and Kazan. According to our assumption, the digital library Lobachevskii DML should be endowed with all conceivable necessary functions and services, making it a comprehensive and up-to-date live DML, generally respected and used by the local as well as the global mathematical community. From the very beginning, we had in mind that the Lobachevskii DML should constitute a building block for the envisioned global WDML.

In this paper, the results of the development of the digital mathematical library Lobachevsky-DML are presented. We describe the purpose of creating this digital library, methods of managing mathematical content based on semantic technologies. The following shows how Lobachevskii DML interacts with the information systems of scientific journals. We also present a system of services to support the life cycle of a mathematical document and highlight technologies for supporting new forms of scientific publications and providing integration services with other digital mathematical archives and libraries.

1 Information Systems in Mathematics

Since inception of the first scientific information systems, the community of mathematicians has been involved in the full cycle of developing such systems, from basic idea to full-scale implementation. Well-known examples are an open source system TeX [6] and commercial systems Wolfram Mathematica and WolframAlpha [7, 8]. With the help of communities of mathematicians, tools for mathematical content management are also actively developed. Examples are MathJax system by American Mathematical Society, information system Math-Net.Ru (http://www.mathnet.ru), developed at the Steklov Mathematical Institute of the Russian Academy of Sciences, and the collection of publicly available preprints arXiv.org (https://arxiv.org/). Now one of the largest digital mathematical libraries is Mizar (http://www.mizar.org/). This is a collection of papers prepared in the Mizar system of formal language, containing definitions, theorems and proofs [9, 10]. Mizar is one of the pioneering systems for mathematics formalization, which still has an active user community. The project has been in constant development since 1973.

At present, scientific research in the field of mathematics is increasingly associated with the use of modern information technologies (cloud, semantic technologies, etc.). These technologies are used in research conducted by distributed scientific groups, the preparation and dissemination of mathematical knowledge in electronic form, the formation of mathematical digital libraries and intellectual processing of their contents. Particular attention is paid to creating a single information space by integrating existing and organizing new digital mathematical libraries (DML). Description of existing digital mathematical libraries with the indication of the purposes and principles of their construction, as well as services for managing scientific content is contained in [11]. Implementation and development of digital mathematical libraries are associated with the development of special tools and the continuous improvement of their functionality.

Since the beginning of the 21st century, a number of developers have created information systems for the management of electronic scientific journals. We compared these systems according to the selected criteria [12]. As a result, the OJS system was recognized as the optimal [13]. We successfully implemented this information platform in the journals Lobachevskii Journal of Mathematics (LJM, http://ljm.kpfu.ru/)/), one of the first Russian electronic mathematical journals, and Russian Digital Libraries (http://ojs.kpfu.ru/index.php/elbib). To manage digital content, we also developed a number of tools that automate a whole series of editorial processes. These include, in particular, the choice of recommendations for the selection of reviewers, style validation of author's documents, search tools, article design services, etc. These tools served as the technological basis for the Lobachevskii DML developed by us (see Section 3).

Let us also pay attention to important results related to the formalization of the of mathematical articles representations. For these purposes, specialized formal languages for the presentation of mathematical texts have been developed (see, for example, [14–18]). These technologies are also used to construct a mathematical ontology and create a semantic search service [19–22].

The above, as well as many other implemented mathematical projects paved the way for the realization of a new idea – the creation of the World Digital Mathematical Library (WDML).

The idea of creating a WDML arose in 2002. The initial aim of this project was digitizing the entire set of mathematical literature (both modern and historical), link it to the present literature, and make it clickable (see [1, 23–28]). As noted in [25], the success of this project and its further impact on mathematics, science and education could be the most significant event after the invention of scientific journals and to become a prototype for a new model of scientific and technical cooperation, a new paradigm for future science in the electronic world. At the same time, the implementation of such a large project will inevitably cause a series of problems. These problems and ways to overcome them were analyzed in [27]. In particular, one of the recommendations was the proposal to develop and coordinate some local projects of creating DML (see ([27, 28]).

Basic plans for the construction of WDML in 2014-2015 were discussed by various mathematical communities and fixed in a number of documents (see [5, 29]). In particular, it was noted that the next step in the development of the WDML project would be building information networks, knowledge-based, which are contained in mathematical publications. Many of the research groups of mathematicians throughout the world took part in the discussion of these ideas, including our group, which represented Kazan University.

In February 2016, in the Fields Institute (Toronto, Ontario) by the Wolfram Foundation, the Fields Institute and the International Mathematical Union working group for the creation WDML, a Seminar on the Semantic Representation of Mathematical Knowledge was organized (https://www.fields.utoronto.ca/Programs/science/15-16/semantic/). Our report on this symposium was devoted to modeling and software solutions in the area of semantic representation of mathematical knowledge [30]. These results correspond to the general ideology of the WDML-project in terms of semantic representation and processing of mathematical knowledge and are a strategic direction of our group's research. In particular, they are connected to the project for the construction of the digital mathematical library Lobachevskii DML, which is described below.

2 Object Approach to the Scientific Digital Content Management

Managing of digital mathematical documents is a unique and complex task. This is due both to the processing of mathematical formulas and to the specific structure of a mathematical document consisting of a logically connected sequence of definitions, theorems, proofs and references. The key idea identified in the WDML project documents is the development of object classes for adequate description and research of mathematical content: a new paradigm for representing digital mathematical content based on elements (classes) and their interrelations is proposed. The selection of classes of mathematical objects and the formation on their basis of ontologies of knowledge areas will allow creating new tools for processing information, in particular, extracting and processing formulas, searching for similar results, and so on.

In our works [31–33], methods of structural analysis of mathematical documents and the selection of objects from them are proposed. In [22, 34], the digital ecosystem OntoMath is described, consisting of ontologies, text analytics tools and applications, designed to control mathematical knowledge. Semantic annotation of mathematical texts is based on the ontology constructed within the framework of the Mocassin project, and ontology of professional mathematics OntoMathPRO [35]. An important application developed on the basis of these ontologies is a special software platform for preparing a mathematical set of related data for publication in the LOD cloud. Another important tool is the semantic search service by mathematical formulas [34]. Another application of the OntoMath ecosystem is a recommendation system for collections of physical and mathematical documents. In particular, for a given document based on selected objects of mathematical knowledge, this system allows you to create a list of "similar" documents (see [36]). These tools are included in the services of the Lobachevskii DML.

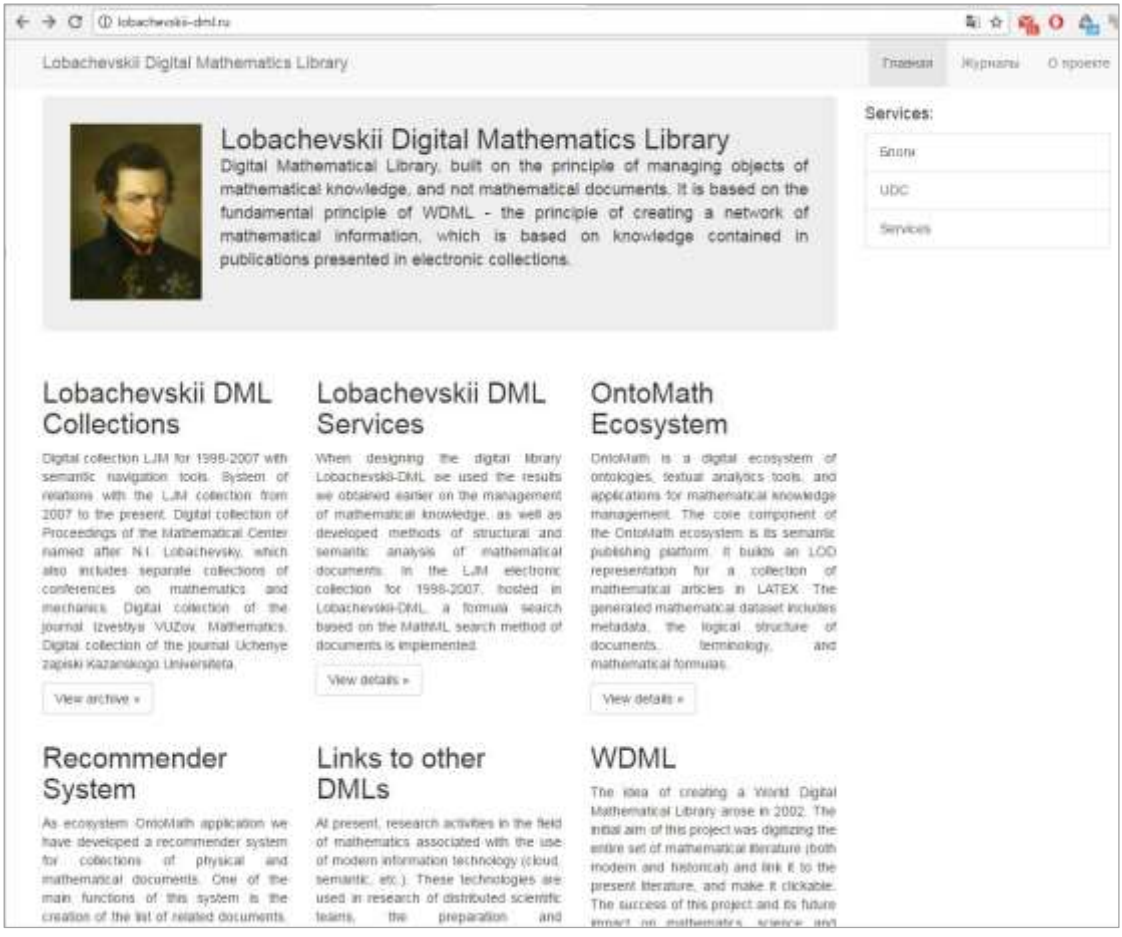


Figure 1 The Lobachevskii DML main page

3 Content, Structure and Services of Lobachevskii DML

3.1 Content and structure of the digital library

Mathematical literature that has been published in Kazan University for more than two centuries of its existence is fairly varied. All these publications have been fully preserved in Scientific Library of Kazan University. The Great Russian scientist N. I. Lobachevskii played a great role in the development of this library. In 1825, he was elected a librarian and managed the library until 1835, combining these duties with the duties of the rector. With him, the scientific foundations of the collection of Library funds were laid, the compilation of single catalogs began, Library has become public, accessible to residents of the city, and a special building was erected for it. Today our library is named after N. I. Lobachevskii. With nearly 6 million publications, it is one of the largest libraries in Russia.

The main goal of creating the Lobachevskii DML was to not only create an archive of specialized literature for mathematics researchers, but also to form an open extensive mathematical library for a wide range of users, possessing a wide range of information processing tools, including search tools. Of course, the basis of the Lobachevskii DML were research journals, as well as selected conference materials and monographs. The main emphasis in the selection of documents was made on the relevance and scientific novelty of materials included in the library.

The informational basis of digital collections in the Lobachevskii DML was printed mathematical books prepared at the Kazan University. At first, these editions were translated into digital format: the result of scanning was a set of pdf-files containing articles from scientific journals, proceedings of conferences, and monographs.

Scientific journals are not only the most important section of the generated DML, but also collections, the easiest to process in DML, because these articles have a uniform structure and a standard set of metadata.

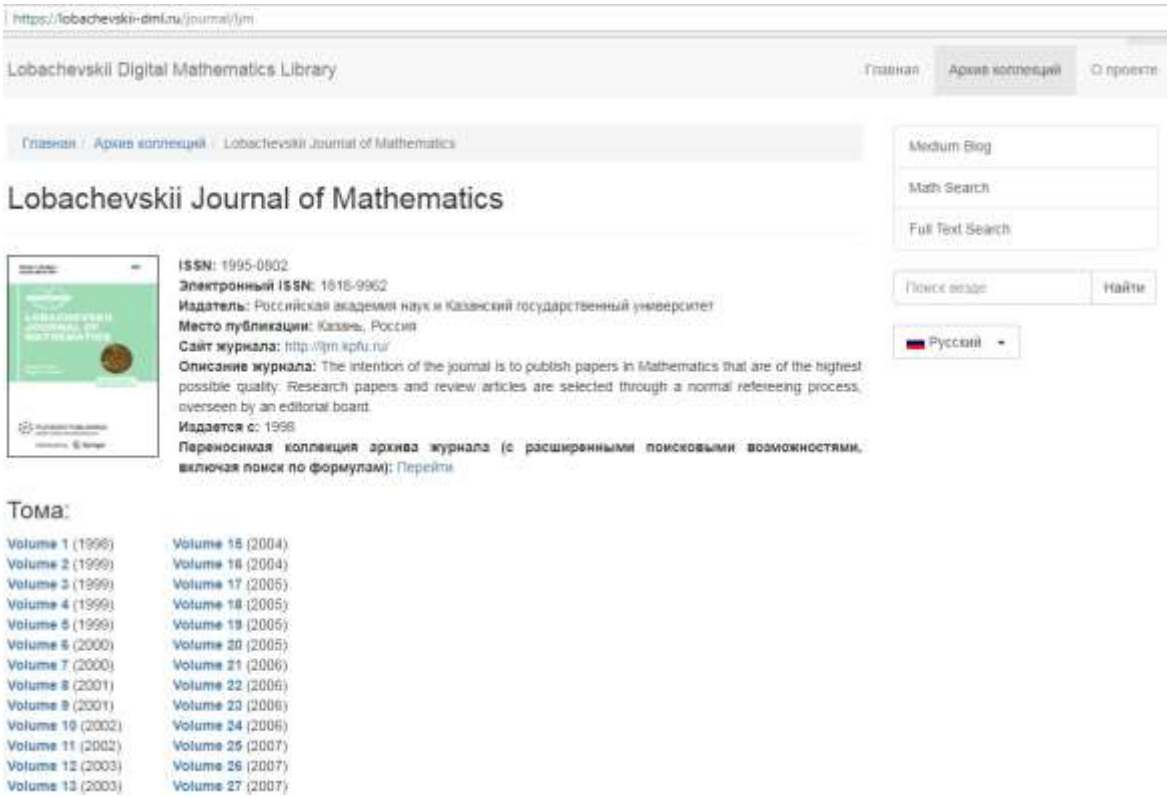


Figure 2 The page of LJM with MathML search tools

Lobachevskii DML structure is as follows:

- the digital collection of articles from the Lobachevskii Journal of Mathematics for 1998–2007 with semantic navigation tools, including search by formulas (see [14]); all the documents of this collection have been translated into the MathML-format (Figure 2);
- the digital collection of LJM articles from 2007 to the present, published by Pleades Publishing LTD and distributed by Springer Science+Business Media LLC;
- the digital collection "Proceedings of the N. I. Lobachevskii Mathematical Center" including materials of international conferences on mathematics and mechanics;
- the digital collection of articles from the journal "Russian Mathematics (Iz. VUZ)";
- the digital collection of articles from the journal "Uchenye Zapiski Kazanskogo Universiteta. Seriya Fiziko-Matematicheskie Nauki (Proceedings of Kazan University. Physics and Mathematics Series).

The digital collections of Lobachevskii DML are shown in Figure 3.

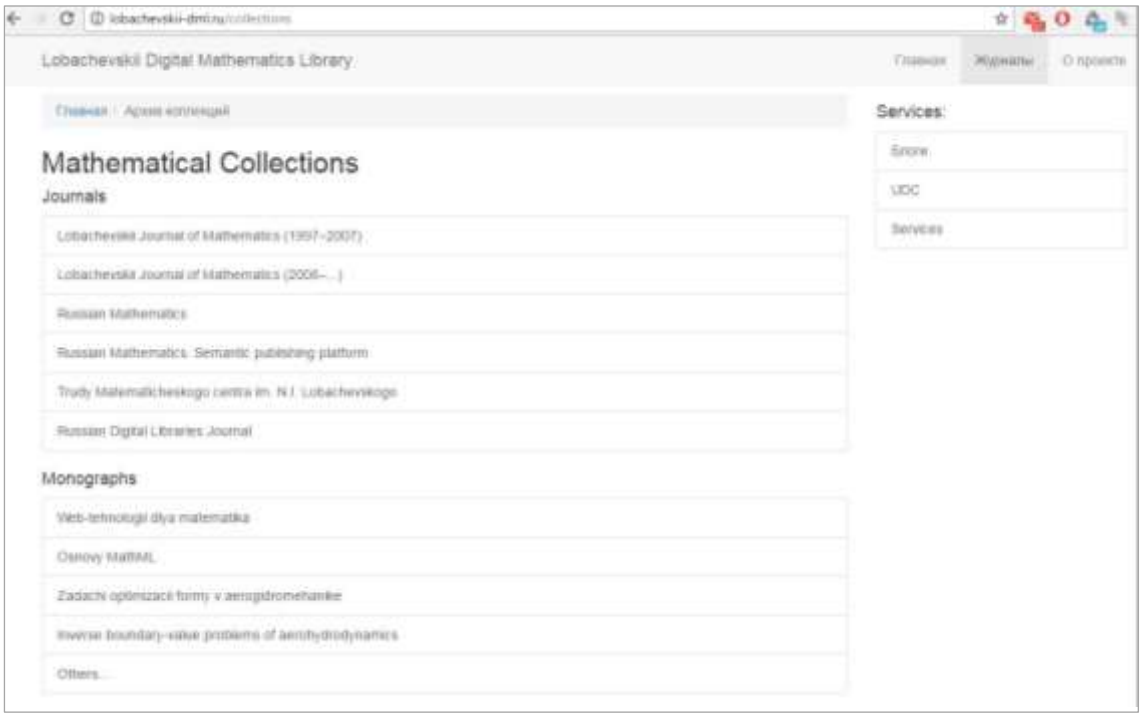


Figure 3 Mathematical collections (https://lobachevskii-dml.ru/collections/)

3.2 Services of Lobachevskii DML

When designing Lobachevskii DML, we used the results we obtained earlier on the management of mathematical knowledge, as well as the developed methods of structural and semantic analysis of mathematical documents [22, 31–36].

For the semantic presentation of documents included in the digital library Lobachevskii DML, we developed an XML-language consisting of a set of tags; the rules for filling them in the form of DTD and XML Schema (see [14]). The XML-file describing each collection of scientific documents was formed in several stages. Each of them assumed the development of a special software tool that eliminates or simplifies the manual processing of a set of digitized documents [31–33, 37, 38]. This XML-file was supplemented with a bibliographic description of the articles contained in the digitized edition (data on the journal number, the title of the conference proceedings, etc.) were added. The most difficult task in this case was the selection of a range of pages for each article included in the publication. Based on the described information, the general file of the processed edition in automatic mode was divided into files containing information about a separate article. Then, in the XML-file that characterizes the collection as a whole, a link to the file describing the article was added.

One of the new search services implemented in Lobachevskii DML is OntoMath Formula Search Engine (http://lobachevskii-dml.ru:8890/mathsearch/). It is a single-page web application that interacts through the SPARQL endpoint with a semantic view of publications using OpenLink Virtuoso (https://virtuoso.openlinksw.com/). This semantic representation is constructed using a semantic publication platform [34]. OntoMath Formula Search finds mathematical formulas containing variables that denote a given mathematical concept. The search is performed in the collection of mathematical documents presented in Lobachevskii DML (Figure 4).

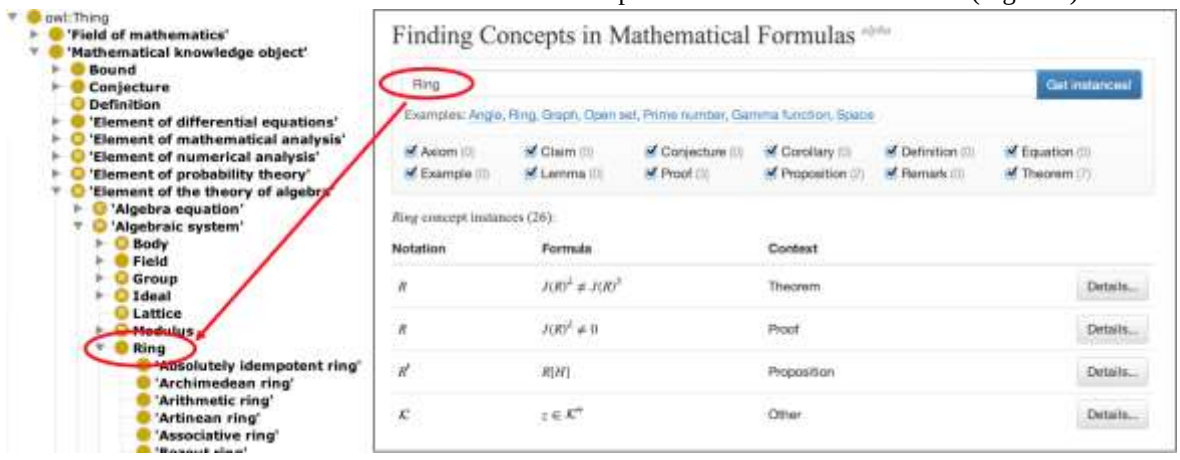


Figure 4 OntoMathPRO Search User Interface

Currently, a new environment of scientific and educational activities is being actively formed, based on the use of Internet technologies. In this regard, along with traditional forms of scientific exchange, focused mainly on printed publications or their electronic versions, new forms of scientific communications are emerging. As the most notable, we note digital presentations as a necessary attribute of reports at scientific conferences, scientific forums and blogs, electronic preprints, webinars and video lectures [39].

The Internet activity of a scientist, often considered as his duty (see, for example, [40]), involves the involvement of all possible means of communication. The use of new forms of scientific exchange should not violate the established traditions of the scientific community, providing for the evaluation of scientific work in the form of peer review, citation system, etc. Consequently, a scientific document of any form should have a bibliographic description and a set of metadata. For example, for "live publications" it is suggested to include in the description the date of the last edition [41]. A successful example of the implementation of the "live publications" model is the Stanford Philosophical Encyclopedia (https://plato.stanford.edu/). All the articles of this encyclopedia were written by specialists in the relevant disciplines and passed the review procedure. The authors maintain the articles up-to-date, periodically updating them. Each new update of the article again undergoes a review procedure, and the history of the publication versions is maintained (it is possible to refer to both the last and any of the previous versions of the article).

Another means of supporting the Internet activity of a scientist are blogs (see, for example, [42, 43]). An example is the blog of modern mathematician Stephen Wolfram (http://blog.stephenwolfram.com/). Another example is the WDML project blog (https://blog.wias-berlin.de/ imu-icm-panel-wdml/). Blogs can be used as a means of organizing open scientific peer review. Such a review, in addition to the traditional one, avoids conflicts of interest and draws a wider circle of experts to the examination. One of the types of open peer review is crowdsourcing-review, in which any representative of the scientific community can take part in the review process.

4 Conclusion

A new digital mathematical library Lobachevskii DML is presented. It is organized based on object management, which corresponds to the paradigm of the World Digital Mathematics Library project. The services for managing mathematical knowledge, implemented in Lobachevskii DML, are described.

This work was funded by the subsidy allocated to Kazan Federal University for the state assignment in the sphere of scientific activities, grant agreement no. 1.2368.2017) and with partial financial support of the Russian Foundation for Basic Research and the Government of the Republic of Tatarstan, within the framework of scientific projects Nos. 15-07-08522, 15-47-02472.

References

[1] Bouche, T.: Digital Mathematics Libraries: The Good, the Bad, the Ugly. Mathematics in Computer Science, 3, pp. 227-241 (2010) doi: 10.1007/ s11786-010-0029-2

[2] Chebukov, D.E., Izaak, A.D., Misyurina, O.G., Pupyrev, Yu.A., Zhizhchenko, A.B.: Math-Net.Ru as a Digital Archive of the Russian Mathematical Knowledge from the XIX Century to Today. Intelligent Computer Mathematics. LNCS, 7961, pp. 344-348 (2013) doi: 10.1007/978-3-642-39320-4_26

[3] Bartošek, M., Lhoták, M., Rákosník, J., Sojka, P., Šárfy, M.: The DML-CZ Project: Objectives and First Steps. Borwein J.M., Rocha E.M., Rodrigues J.F. (eds.) Communicating Mathematics in the Digital Era, pp. 75-86. A K Peters, Ltd. (2008)

[4] Bartošek, M., Rákosník, J.: DML-CZ: The Experience of a Medium-Sized Digital Mathematics Library. Notices of the AMS, 60 (8), pp. 1028-1033 (2013) doi: http://dx.doi.org/10.1090/ noti1031

[5] Developing a 21st Century Global Library for Mathematics Research. Washington, The National Academies Press (2014). doi: 10.17226/18619

[6] Knuth, D.E.: The TeX Book. Addison-Wesley Publishing Company (1986)

[7] Wolfram, S.: A New Kind of Science. Wolfram Media, Inc. (2002)

[8] Wolfram, S.: An Elementary Introduction to the Wolfram Language. Wolfram Media, Inc. (2015)

[9] Naumowicz, A., Kornilowicz, A.: A Brief Overview of Mizar. S. Berghofer et al. (Eds.), TPHOLs 2009, LNCS 5674, pp. 67-72, Springer-Verlag (2009)

[10] Bancerek, G., Bylinski, C., Grabowski, A., Kornilowicz, A., Matuszewski, R., Naumowicz, A., Pak, K., Urban, J.: Mizar: State-of-the-Art and Beyond. M. Kerber et al. (Eds.), Intelligent Computer Mathematics, CICM 2015, LNAI 9150, pp. 261-279 (2015)

[11] Elizarov, A.M., Lipachev, E.K., Zuev, D.S.: Digital Mathematical Libraries: Overview of Implementations and Content Management Services. Current Proceedings

[12] Elizarov, A.M., Lipachev, E.K., Zuev, D.S.: Infrastructure of Electronic Scientific Journal and Cloud Services Supporting Lifecycle of Electronic Publications. CEUR Workshop Proceedings, 1297, pp. 156-159 (2014), http://ceur-ws.org/Vol-1297/ 156-159_paper-23.pdf

[13] MacGregor, J., Stranack, K., Willinsky, J.: The Public Knowledge Project: Open Source Tools for

Open Access to Scholarly Communication. Bartling S., Friesike S. (Eds) Opening Science. The Evolving Guide on How the Internet is Changing Research, Collaboration and Scholarly Publishing. Springer International Publishing, pp. 165-175 (2014) doi:10.1007/978-3-319-00026-8_11

[14] Elizarov, A.M., Lipachev, E.K., Malakhaltsev, M.A.: Web Technologies for Mathematicians: The Basics of MathML. A Practical Guide. Moscow: Fizmatlit, 192 p. (2010) (in Russian)

[15] Kohlhase, M.: An Open Markup Format for Mathematical Documents (Version 1.2). LNAI 4180. Springer Verlag (2006). http://omdoc.org/pubs/ omdoc1.2.pdf

[16] Iancu, M., Kohlhase, M., Rabe, F., Urban, J.: The Mizar Mathematical Library in OMDoc: Translation and Applications. Journal of Automated Reasoning, 50 (2), pp. 191-202, Springer Verlag (2013)

[17] Kohlhase, M.: Semantic Markup in TeX/ LaTeX. http://ctan.altspu.ru/macros/latex/contrib/stex/sty/ stex/stex.pdf

[18] Dehaye, P., Iancu, M., Kohlhase, M., Konovalov, A., Lelièvre, S., Müller, D., Pfeiffer, M., Rabe, F., Thiéry, N.M., Wiesing, T.: Interoperability in the OpenDreamKit Project: the Math-in-the-middle Approach. Intelligent Computer Mathematics, M. Kohlhase, M. Johansson, B. Miller, L. de Moura, F. Tompa (Eds.), LNCS, 9791, pp. 117-131 (2016). https://github.com/ OpenDreamKit/OpenDreamKit/blob/master/WP6/ CICM2016/published.pdf

[19] Lange, C.: Ontologies and Languages for Representing Mathematical Knowledge on the Semantic Web. Semantic Web, 4 (2), pp. 119-158 (2013), doi: 10.3233/SW-2012-0059

[20] Lange, C.: Enabling Collaboration on Semiformal Mathematical Knowledge by Semantic Web Integration. Ph.D. Thesis, Jacobs University Bremen (2011)

[21] Elizarov, A., Kirillovich, A., Lipachev, E., Nevzorova, O., Solovyev, V., Zhiltsov N.: Mathematical Knowledge Representation: Semantic Models and Formalisms. Lobachevskii J. of Mathematics, 35 (4), pp. 347-353 (2014), doi: 10.1134/ S1995080214040143

[22] Elizarov, A.M., Kirilovich, A.V., Lipachev, E.K., Nevzorova, O.A.: Mathematical Knowledge Management: Ontological Models and Digital Technology. CEUR Workshop Proceedings, 1752, pp. 44-50 (2016), http://ceur-ws.org/Vol-1752/ paper08.pdf

[23] Jackson, A.: The Digital Mathematics Library. Notices of the AMS, 50 (4), pp. 918-923 (2003). http://www.ams.org/notices/200308/comm-jackson.pdf

[24] The Digital Mathematical Library Project. Status August 2005. http://www.math.uiuc.edu/~tondeur/ DML04.pdf

[25] Digital Mathematics Library: a Vision for the Future. International Mathematical Union (2006). http://www.mathunion.org/fileadmin/IMU/Report/ dml_vision.pdf

[26] Tondeur, P.: WDML: The World Digital Mathematics Library. The Evolution of Mathematical Communication in the Age of Digital Libraries. IMA Workshop, December 8–9, 2006. http://www.math.uiuc.edu/~tondeur/WDML_ IMA_DEC2006.pdf

[27] Sylwestrzak, W., Borbinha, J., Bouche, T., Nowinski, A., Sojka, P.: EuDML – Towards the European Digital Mathematics Library. P. Sojka (ed.) Towards a Digital Mathematics Library. Paris, July 7–8th, 2010, pp. 11-26. Masaryk University Press, Brno (2010). http://dml.cz/bitstream/ handle/10338.dmlcz/702569/DML_003-2010-1_5.pdf

[28] Pitman, J., Lynch, C.: Planning a 21st Century Global Library for Mathematics Research. Notices of the AMS, 61 (7), pp. 776-777 (2014). http:// www.ams.org/notices/201407/rnoti-p776.pdf

[29] Olver, P.J.: The World Digital Mathematics Library: Report of a Panel Discussion. Proceedings of the International Congress of Mathematicians, August 13–21, 2014, Seoul, Korea. Kyung Moon SA, 1, pp. 773-785 (2014)

[30] Elizarov, A.M., Zhiltsov, N.G., Kirillovich, A.V., Lipachev, E.K., Nevzorova, O.A., Solovyev, V.D.: The OntoMath Ecosystem: Ontologies and Applications for Math Knowledge Management. Semantic Representation of Mathematical Knowledge Workshop 5 February 2016. http://www.fields. utoronto.ca/video-archive/ 2016/02/ 2053–14698

[31] Elizarov, A.M., Lipachev, E.K., Hohlov, Yu.E.: Semantic Methods of Structuring Mathematical Content Providing Enhanced Search Functionality. Information Society, 1–2, pp. 83-92 (2013), http://elibrary.ru/download/elibrary_20376784_ 48362557.pdf

[32] Biryal'tsev, E., Elizarov, A., Zhil'tsov, N., Lipachev, E., Nevzorova O., Solov'ev, V.: Methods for Analyzing Semantic Data of Electronic Collections in Mathematics. Automatic Documentation and Mathematical Linguistics, Allerton Press, Inc. 48 (2), pp. 81-85 (2014). doi: 10.3103/ S000510551402006X

[33] Elizarov, A., Lipachev, E., Nevzorova O., Solov'ev, V.: Methods and Means for Semantic Structuring of Electronic Mathematical Documents. Doklady Mathematics, 90 (1), pp. 521-524 (2014). doi:10.1134/S1064562414050275

[34] Elizarov, A., Kirillovich, A., Lipachev, E., Nevzorova, O.: Digital Ecosystem OntoMath:

Mathematical Knowledge Analytics and Management. Communications in Computer and Information Science, Springer, 706, pp. 33-46 (2017). doi: 10.1007/978-3-319-57135-5_3

[35] Elizarov, A., Zhil'tsov, N., Kirilovich, A., Lipachev, E.: Semantic Annotation in the Control System of Physical and Mathematical Content. Scientific Service in the Internet: Works of the XVII All-Russian Scientific Conference. Moscow: M. V. Keldysh Institute of Applied Mathematics, pp. 98-103 (2015)

[36] Elizarov, A.M., Kirillovich, A.V., Lipachev, E.K., Zhizhchenko, A.B., Zhil'tsov, N.G.: Mathematical Knowledge Ontologies and Recommender Systems for Collections of Documents in Physics and Mathematics. Doklady Mathematics, 93 (2), pp. 231-233 (2016), doi:10.1134/ S1064562416020174

[37] Elizarov, A., Zuev, D., Lipachev, E., Malakhaltsev, M.: Services Structuring Mathematical Content and Integration of Digital Mathematical Collections into Scientific Information Space. CEUR Workshop Proceedings, 934, pp. 309-312 (2012). http://ceur-ws.org/Vol-934/ paper47.pdf

[38] Elizarov, A.M., Lipachev, E.K., Haidarov, S.M.: Automated Processing Service System of Large Collections of Scientific Documents. CEUR Workshop Proceedings, 1752, pp. 58-64 (2016), http://ceur-ws.org/Vol-1752/paper10.pdf

[39] Chebukov, D., Izaak, A., Misyurina, O., Pupyrev, Yu.: Math-Net.Ru Video Library: Creating a Collection of Scientific Talks. Mathematical Software – ICMS 2016, 5th Int. Conference, Berlin, Germany, July 11–14, 2016, Proceedings, Theoretical Computer Science and General Issues, LNCS, 9725, eds. G.-M. Greuel, Th. Koch, P. Paule, A. Sommese, Springer, pp. 447-450 (2016)

[40] Gorbunov-Posadov, M.M.: Internet Activity as a Scientist's Duty. Revision of 25.02.2017. http:// keldysh.ru/gorbunov/duty.htm

[41] Gorbunov-Posadov, M.M., Skornyakova, R.Yu.: The Date of the Last Edition as a Living Attribute of a Live Publication. Scientific Service in the Internet: Works of the XVIII All-Russian Scientific Conference (September 19–24, 2016, Novorossiysk). Moscow: M. V. Keldysh Institute of Applied Mathematics, pp. 113-114 (2016). doi: 10.20948/abrau-2016-48

[42] Puschmann, C.: (Micro)Blogging Science? Notes on Potentials and Constraints of New Forms of Scholarly Communication. S. Bartling, S. Friesike (Eds) Opening Science. The Evolving Guide on How the Internet is Changing Research, Collaboration and Scholarly Publishing. Springer International Publishing, pp. 89-106 (2014). doi: 10.1007/978-3-319-00026-8_6

[43] Elizarov, A.M., Kirilovich, A.V., Lipachev, E.K.: Blogs in Scientific Communications Systems. Scientific Notes of the Institute of Social and Humanitarian Knowledge. Kazan: Institute of Social and Humanitarian Knowledge, 1 (15), pp. 209-214 (2017)

Представление и извлечение знаний

Knowledge representation and discovery

Семантическая модель представления и обработки баз знаний

© В.В. Голенков © Н.А. Гулякина © И.Т. Давыденко © Д.В. Шункевич

Белорусский государственный университет информатики и радиоэлектроники,
Минск, Беларусь

golen@bsuir.by guliakina@bsuir.by davydenko@bsuir.by shunkevichdv@gmail.com

Аннотация. Предложен подход к созданию интеллектуальных систем, ориентированных на решение комплексных задач, в основе которого лежат семантические модели баз знаний и согласованные с ними семантические модели машин обработки базы знаний. Основой для построения указанных моделей является унифицированное смысловое представление знаний на основе универсального языка семантических сетей с теоретико-множественной интерпретацией. На базе указанного языка построено открытое семейство совместимых языков, семантика каждого из которых задается соответствующей онтологией. Семантическая модель машины обработки базы знаний построена на базе многоагентного подхода, предполагающего, что агенты взаимодействуют между собой через общую для них семантическую память.

Ключевые слова: база знаний, машина обработки базы знаний, модели представления баз знаний, модели обработки баз знаний, семантические модели, семантические сети, семантическая память, предметные области, онтологии, многоагентные системы над общей памятью.

Semantic Model of Knowledge Bases Representation and Processing

© V.V. Golenkov © N.A. Guliakina © I.T. Davydenko © D.V. Shunkevich

Belarusian State University of Informatics and Radioelectronics,
Minsk, Belarus

golen@bsuir.by guliakina@bsuir.by davydenko@bsuir.by shunkevichdv@gmail.com

Abstract. The article offers an approach to the creation of intelligent systems based on semantic models of knowledge bases and compatible semantic models of knowledge base processing machines. Such intelligent systems are intended to solve various difficult and complex problems. Unified semantic representation of knowledge forms the basis for the aforementioned models and is itself based on the universal language of semantic networks with set-theoretic interpretation. This language is used to build an open family of compatible languages. Their semantics are specified by the corresponding ontologies. Multi-agent approach is used to build the semantic model of the knowledge base processing machine. This approach assumes that agents interact with each other through a shared semantic memory.

Keywords: knowledge bases, knowledge base processing machine, knowledge bases representation models, knowledge bases processing models, semantic models, semantic networks, semantic memory, subject areas, ontologies, multiagent systems over shared memory.

1 Введение

Современные интеллектуальные системы представляют собой естественный этап эволюции компьютерных систем и, в частности, эволюции моделей представления информации в памяти компьютерных систем, а также моделей обработки информации.

Эволюция представления обрабатываемой информации в памяти компьютерных систем осуществляется в различных направлениях: от неструктурированных данных (наборов значений заданных параметров) к структурированным данным (матрицам, списковым структурам, реляционным структурам); от простых реляционных структур к реляционным структурам со связями, компонентами которых являются другие связи или целые подструктуры; от данных к метаданным; от описания постоянных (постоянно существующих) сущностей к описанию временных сущностей, к описанию их прошлого, настоящего и будущего; от описания стационарных сущностей к описанию нестационарных сущностей, у которых меняется их

Труды XIX Международной конференции «Аналитика и управление данными в областях с интенсивным использованием данных» (DAMDID/ RCDL'2017), Москва, Россия, 10–13 октября 2017 года

состояние и внутренняя структура; от данных, структуризация которых определяется исключительно «интересами» использующей их программы, к данным, структуризация которых определяется их смыслом, и, следовательно, обработка которых может осуществляться с помощью произвольного набора программ; от фактографических высказываний к логическим (с переменными, логическими связками и кванторами); от данных, семантическая структуризация которых явно не задана, к знаниям, в которых явно выделены знания различного вида (предметные области, онтологии) и явно описаны связи между ними; от неявно формулируемых задач к явным формулировкам задач; от четких, точных, достоверных знаний к нечетким, неточным, правдоподобным знаниям.

Эволюция моделей обработки информации в памяти компьютерных систем прежде всего определяется эволюцией операционной семантики языков программирования: от последовательных программ (в частности, алгоритмов) к параллельным (синхронным и асинхронным) программам; от процедурных (императивных) программ к непроцедурным (функциональным, логическим) программам; от «жестких» вычислений к «мягким» (нечетким логическим программам, генетическим алгоритмам, нейронным сетям); от программ, доминирующих над данными, к программам, в которых доминируют обрабатываемые ими данные, структурируемые независимо от программ, использующих эти данные; от «пассивных» программ, инициируемых извне, к активным, самоинициируемым, агентным программам.

Основными компонентами интеллектуальной системы являются ее база знаний, включающая в себя всю информацию, которую интеллектуальная система использует в процессе своего функционирования, а также машина обработки указанной базы знаний, включающая в себя все функциональные возможности заданной интеллектуальной системы.

Расширение областей применения интеллектуальных систем требует поддержки решения комплексных задач, каждая из которых предполагает согласованное применение различных моделей представления и различных моделей обработки знаний.

Для решения комплексных задач требуется обеспечить совместимость и интеграцию самых различных моделей представления знаний и моделей их обработки [1]. Многообразие видов знаний, используемых интеллектуальными системами, многообразие формальных моделей представления этих знаний, формальных моделей решения задач, моделей обработки знаний необходимо превратить в новое качество, предполагающее согласованное использование всех этих моделей, т. е. предполагающее интеграцию самых разнообразных информационных ресурсов и сервисов [2].

Решение проблемы совместимости и одновременного использования (в ходе решения одной задачи) различных моделей представления знаний и различных моделей обработки знаний, различных моделей решения задач означает переход к принципиально новому этапу эволюции компьютерных систем – интеллектуальным системам нового поколения, ориентированным на решение комплексных задач. Каждая такая система характеризуется следующими особенностями:

- Вся информация, хранимая в памяти компьютерной системы, систематизирована в виде единой базы знаний (т. е. любой фрагмент информации входит в состав базы знаний). К такой информации относятся непосредственно обрабатываемые знания, интерпретируемые программы, формулировки решаемых задач, планы и протоколы решения задач, информация о пользователях, описание синтаксиса и семантики внешних языков, описание пользовательского интерфейса и многое другое.
- Обеспечивается совместимость всех видов знаний, используемых в компьютерной системе.
- Вся обработка информации ориентирована на обработку целостной хорошо структурированной базы знаний и управляется этой базой знаний.
- Обеспечивается совместимость всевозможных моделей обработки информации и всевозможных моделей решения задач.
- Обеспечивается поддержка высоких темпов эволюции интеллектуальных систем в ходе их эксплуатации.
- Обеспечивается поддержка высоких темпов эволюции самой технологии разработки интеллектуальных систем.

На создание интеллектуальных систем такого рода ориентирована Технология OSTIS, разработка которой ведется авторами данной статьи. Частные результаты, полученные по данной тематике, опубликованы в ряде работ авторов, например, [3–5].

Целью данной работы является систематизация указанных частных результатов и рассмотрение связи между моделями представления знаний и моделями обработки знаний (модель обработки знаний должна учитывать то, как эти знания устроены)

2 Предлагаемый подход

Основой предлагаемого подхода к построению баз знаний и машин обработки баз знаний в интеллектуальных системах, ориентированных на решение комплексных задач, являются *семантические модели баз знаний* и согласованные с ними *семантические модели машин обработки баз знаний.*

В основе понятия *семантической модели базы знаний* лежат следующие положения:

- Внутреннее представление базы знаний в памяти интеллектуальной системы осуществляется в форме смыслового представления в виде формализованной семантической сети.

- В рамках базы знаний осуществляется явное выделение предметных областей и явное представление онтологий, описывающих семантику всех рассматриваемых в базе знаний предметных областей и соответствующих им языков.
- Осуществляется онтологическая структуризация базы знаний в виде иерархической системы предметных областей и соответствующих им онтологий.
- Используется широкий набор видов структуризации базы знаний.

В основе понятия *семантической модели машины обработки знаний* лежат следующие положения:

- Рассматривается обработка баз знаний, представленных в виде их семантических моделей.
- Вводится понятие абстрактной семантической памяти, которая трактуется как динамическая среда, в каждый момент времени отражающая текущее состояние семантической модели обрабатываемой базы знаний. Процесс обработки базы знаний, которая хранится в семантической памяти в виде семантической сети, сводится не только к изменению состояния элементов этой семантической сети, но и к изменению конфигурации связей между указанными элементами (к удалению одних связей и генерации других).
- Вводится предметная область, объектами исследования которой являются целенаправленные процессы и соответствующие им задачи, решаемые в рамках семантической памяти.
- Вводится предметная область, объектами исследования которой являются агенты, выполняющие указанные процессы в этой памяти.
- Строятся и явно включаются в состав обрабатываемой базы знаний онтологии, описывающие семантику (спецификацию понятий) предметной области целенаправленных процессов и задач, решаемых в семантической памяти, и предметной области агентов, выполняющих эти процессы.

Для того чтобы превратить различного вида знания, хранимые в памяти компьютерной системы, в единую, хорошо структурированную базу знаний, необходимо:

- привести все эти разнообразные виды знаний к единому синтаксическому и семантическому фундаменту, основанному на некоторой универсальной онтологии представления [6, 7];
- разработать типологию сущностей, описываемых в базе знаний, а также семейство онтологий, соответствующих основным типам сущностей;
- разработать такую типологию знаков, входящих в состав базы знаний, которая отражает не типологию обозначаемых ими сущностей, а характер соотношения указанных знаков с текущим состоянием базы знаний, отражающего степень полноты сведений об обозначаемых сущностях;
- разработать предметную область и онтологию всевозможного вида знаний, хранимых в составе базы знаний, которые рассматривают знания как важнейший вид сущностей, описываемых в базе знаний, и в которых исследуются типология знаний, отношения, заданные на знаниях.
- обеспечить возможность неограниченного перехода от знаний к соответствующим им метазнаниям.

Для внутреннего представления знаний в памяти компьютерной системы нами предлагается открытое семейство совместимых языков, каждый из которых является подъязыком базового языка смыслового представления знаний, рассматриваемого ниже, и семантика каждого из которых описывается соответствующей онтологий.

3 Принципы внутреннего смыслового представления знаний

Основное требование, предъявляемое к ***формальному языку смыслового представления знаний***, – это устранение семантической эквивалентности текстов в рамках базы знаний каждой интеллектуальной системы. Таким образом, смысловое представление знания можно трактовать как инвариант многообразия семантических форм представления этого знания.

В качестве базового внутреннего формального языка представления знаний в памяти интеллектуальных систем предлагается язык, названный нами ***SC-кодом*** (Semantic Computer Code). С формальной точки зрения ***SC-код*** есть множество текстов (sc-текстов), теоретико-множественной объединение которых представляет собой бесконечную структуру, включающую в себя описания всевозможных сущностей.

Все синтаксически элементарные (атомарные) фрагменты текстов SC-кода являются ***знаками*** соответствующих им (обозначаемых ими) *сущностей*. Такие элементарные фрагменты sc-текстов будем называть ***sc-элементами***.

С формальной точки зрения SC-код является языком семантических сетей. Основное достоинство семантических сетей и текстов SC-кода в частности – это соединение синтаксического и семантического аспектов представления знаний, что значительно снижает вычислительную сложность обработки знаний [8].

Подчеркнем, что переход от традиционных текстов к семантическим сетям можно рассматривать как процесс избавления от тех языковых излишеств, которые обусловлены коммуникативной функцией традиционных языков, но не являются необходимыми для построения формальной смысловой внутренней модели мира. Избавление от

указанных излишеств включает в себя: исключение семантически неинтерпретируемых фрагментов текста – букв, разделителей, ограничителей, слов, которые не являются знаками сущностей; исключение синонимии знаков; исключение омонимии знаков.

4 Типология описываемых сущностей и их знаков

Классификация sc-элементов может осуществляться в нескольких аспектах – с точки зрения синтаксической типологии самих знаков, с точки зрения типологии сущностей, обозначаемых этими знаками (семантический аспект); с точки зрения характера соотношения sc-элемента с обозначаемой им сущностью; с точки зрения характера соотношения sc-элемента с присутствующими в текущем состоянии базы знаний сведениями о сущности, обозначаемой этим sc-элементом.

По синтаксическому типу множество sc-элементов разбивается на *sc-узлы* и *sc-коннекторы* (*sc-дуги* – знаки ориентированных бинарных связей; *sc-ребра* – знаки неориентированных бинарных связей).

По признаку константности-переменности множество sc-элементов разбивается на *sc-константы* (константные sc-элементы) и *sc-переменные* (переменные sc-элементы). Тип sc-переменной определяется областью ее возможных значений.

По структурному признаку множество sc-элементов разбивается на *знаки внешних сущностей, знаки множеств sc-элементов* и *знаки терминальных абстрактных сущностей* (т. е. абстрактных сущностей, не являющихся множествами).

В свою очередь, множество знаков множеств sc-элементов по структурному признаку разбивается на *sc-классы* – знаки классов sc-элементов, *sc-связки* – знаки связей между sc-элементами, каждая из которых трактуется как множество связываемых ею sc-элементов, *sc-структуры* – знаки структур, состоящих из sc-элементов в общем случае разного структурного типа.

Каждая ***sc-структура*** представляет собой множество *sc-элементов*, удаление одного из которых может привести к нарушению целостности этого множества. В рамках каждой *sc-структуры* явно указываются роли ее элементов. Более подробно типология sc-структур и средства их спецификации рассмотрены в работе [4].

По **темпоральному признаку** множество sc-элементов разбивается на *знаки постоянных сущностей* и *знаки временных сущностей*.

Более подробно типология сущностей и их знаков рассмотрена в работе [3].

5 Семантическое многообразие и типология знаний

В рамках базы знаний будем выделять семантически осмысленные *sc-структуры,* обладающие некоторой семантической целостностью. Такие структуры будем называть ***знаниями***.

В рамках предлагаемого подхода выделяются такие виды знаний, как *семантическая окрестность*, *предметная область* [9], *онтология*, *раздел базы знаний*, *утверждение*, *определение*, *задача*, *программа*, *план*, *решение*, *сравнение*, *фактографическое знание* и др.

Важнейшим отношением, заданным на множестве знаний, является отношение *быть метазнанием**, описывающее переход от знаний к описывающим их метазнаниям [7]. Связки указанного отношения связывают некоторое исходное *знание* со *знанием*, которое является его спецификацией.

Примером связи между знанием и соответствующим ему *метазнанием** является переход от некоторого исходного знания к описанию его декомпозиции (сегментации) на некоторые части с указанием связей между этими частями.

Более подробно типология знаний и средства их спецификации рассмотрены в работе [4].

6 Семантические окрестности и их типология

Каждая ***семантическая окрестность*** – это знание, являющееся спецификацией (описанием) некоторой *сущности*, знак которой считается *ключевым элементом* этой спецификации.

Выделяются следующие виды семантических окрестностей: *семантическая окрестность по инцидентным sc-коннекторам* (с дополнительным указанием бинарных отношений, которым эти коннекторы принадлежат), *семантическая окрестность по выходящим sc-дугам*, *семантическая окрестность по входящим sc-дугам*, *семантическая окрестность по инцидентным небинарным связкам* (с указанием небинарных отношений, которым эти связки принадлежат), *полная семантическая окрестность*, структура которой определяется семантическим типом специфицируемой сущности, *типовая семантическая окрестность* (минимально достаточная), структура которой также определяется семантическим типом специфицируемой сущности, *определение*, *пояснение*, *примечание*, *правило идентификации экземпляров*, *терминологическая спецификация*, *теоретико-множественная спецификация*, *логическая спецификация*, *описание типичного экземпляра*, *обоснование, структуризация, параметрическая спецификация темпоральная спецификация, пространственная спецификация*.

7 Уточнение понятия предметной области и онтологии

Формальная модель ***предметной области***, представленная в *SC-коде*, является *sc-структурой*, в рамках которой с помощью специального набора

ролевых отношений выделяется ряд *ключевых элементов* этой структуры и указываются их роли в рамках этой структуры. Такие ролевые отношения являются подмножествами *Отношения принадлежности*.

Ключевыми элементами предметной области являются, прежде всего, знаки *рассматриваемых понятий* (концептов), уточнение смысла которых является существенным для семантического анализа указанной предметной области.

Для понятий, которые рассматриваются в предметных областях, возможны все четыре варианта уточнения их ролей: понятие может быть и исследуемым, и вводимым в данной предметной области; понятие может быть исследуемым в данной предметной области, но введенным в другой предметной области; понятие может быть неисследуемым, но вводимым в данной предметной области; понятие может быть и неисследуемым, и невводимым в данной предметной области.

На Рис. 1 на примере фрагментов структурных спецификаций нескольких предметных областей, выделяемых в рамках геометрии Евклида, показан принцип соотнесения понятий с предметными областями. Указанный фрагмент базы знаний представлен на языке SCg [10], который является графическим эквивалентом SC-кода.

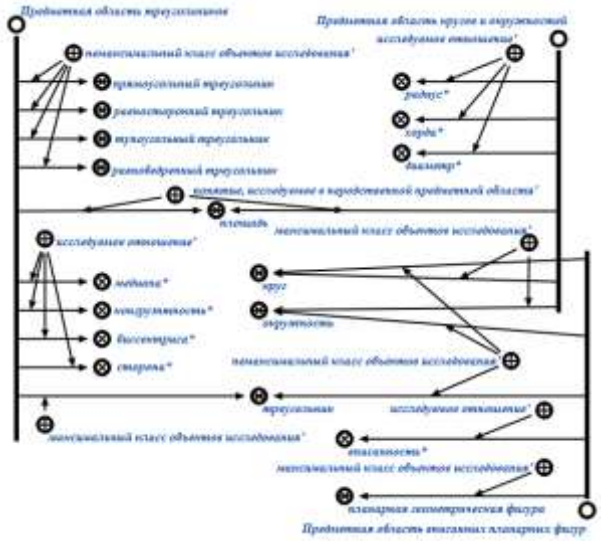


Рисунок 1 Спецификация предметных областей

Описание семантики ключевых понятий предметной области есть не что иное как ***онтология***, соответствующая указанной *предметной области*.

В рамках предлагаемого подхода выделены такие типы онтологий, как *структурная спецификация*, *теоретико-множественная онтология*, *логическая иерархия понятий*, *логическая онтология*, *логическая иерархия высказываний*, *терминологическая онтология*, *онтология задач и решений задач*, *онтология классов задач и способов решения задач*.

Онтология, являющаяся результатом объединения всех онтологий, специфицирующих заданную предметную область, названа *интегрированной онтологией*.

Связь между предметной областью и ее онтологией задается отношением *онтология**, которое является частным видом отношения *метазнание**.

Более подробно понятие предметной области рассматривается в работах [3, 4].

8 Структуризация баз знаний

В рамках предлагаемого подхода структуризация базы знаний может проводиться по различным критериям:

- структуризация базы знаний, отражающая многообразие видов знаний, входящих в ее состав;
- структуризация базы знаний, основанная на иерархии предметных областей и соответствующих им (специфицирующих их) онтологий;
- структуризация базы знаний, отражающая архитектуру интеллектуальной системы (все ее подсистемы, базы знаний, машины обработки базы знаний, пользовательские интерфейсы этих подсистем и базы знаний и машины обработки базы знаний всех указанных выше пользовательских интерфейсов);
- структуризация базы знаний, отражающая динамику самой базы знаний, т. е. внутреннего мир этой базы знаний – истории ее эволюции, ее текущего согласованного (утвержденного) состояния, планов ее совершенствования;
- структуризация базы знаний, отражающая темпоральные свойства внешнего описываемого мира (динамика внешнего мира) – прошлое, настоящее, будущее;
- прагматически ориентированная структуризация базы знаний (для разработчиков и конечных пользователей) – декомпозиция на разделы, отражающие распределение областей доступа для различных категорий пользователей и разработчиков по просмотру и редактированию (текущее состояние согласованной части БЗ, персональные черновики);
- структуризация базы знаний, отражающая авторство различных фрагментов баз знаний, множественных точек зрения (возможно, противоречащих друг другу) и непротиворечивую часть базы знаний, отражающую согласованную точку зрения коллектива авторов данной базы знаний;
- структуризация базы знаний, отражающая соответствие между внешними универсальными и специализированными языками и семантически эквивалентными им внутренними языками (sc-языками и соответствующими предметными областями).

9 Уточнение понятия целенаправленного процесса в семантической памяти

Машина обработки базы знаний оперирует знаниями определенного вида. Важнейшими видами таких знаний являются *процессы*, выполняемые такой машиной, *задачи* (спецификации процессов),

спецификации агентов обработки базы знаний, в том числе, различного рода *программы*, описывающие алгоритмы действий этих агентов. Ниже подробнее рассмотрим принципы формализации перечисленных видов знаний.

В рамках предлагаемого подхода формальная модель некоторого ***процесса*** представляет собой ситуативную *sc-структуру*, в каждый момент времени описывающую текущее состояние объектов, участвующих в данном процессе. Процесс, описывающий изменения, происходящие исключительно в рамках семантической памяти (sc-памяти), будем называть *процессом в sc-памяти*.

Целенаправленный процесс, выполняемый некоторым *субъектом*, будет называть ***действием***.

По отношению к памяти компьютерной системы выделяются такие классы действий, как *информационное действие* (действие в памяти компьютерной системы), *поведенческое действие* (действие во внешней среде), *эффекторное действие*, *рецепторное действие*.

По отношению к текущему моменту времени выделяются такие классы действий, как *инициированное действие*, *планируемое действие*, *выполненное действие*.

Более подробно типология действий и средства их спецификации рассмотрены в работе [5].

В процессе описания в семантической памяти деятельности некоторого коллектива субъектов возникает необходимость выделять в рамках этой деятельности обособленные логически целостные фрагменты, которые могут выполняться отдельными субъектами независимо друг от друга. Классы таких действий названы ***классами логически атомарных действий.***

Каждое *действие*, принадлежащее некоторому конкретному ***классу логически атомарных действий***, обладает двумя необходимыми свойствами:

- выполнение действия не зависит от того, является ли указанное действие частью декомпозиции более общего действия. При выполнении данного действия также не должен учитываться тот факт, что данное действие предшествует каким-либо другим действиям или следует за ними;
- указанное действие должно представлять собой логически целостный акт преобразования семантической памяти. Такое действие, по сути, является транзакцией, т.е. результатом такого преобразования становится новое состояние преобразуемой системы, а выполняемое действие должно быть либо выполнено полностью, либо не выполнено совсем, частичное выполнение не допускается.

10 Уточнение понятия задачи, решаемой в семантической памяти, и многообразие видов задач

В рамках предлагаемого подхода ***задача*** представляет собой спецификацию некоторого действия.

Формулировка каждой *задачи* может включать факт принадлежности *действия* какому-либо частному классу *действий*; описание *цели** (*результата**) *действия*, если она точно известна; указание *заказчика* действия*; указание *исполнителя* действия* (в том числе, коллективного); указание *аргумента(ов) действия'*; указание инструмента или посредника *действия*; описание *декомпозиции действия**; указание *последовательности действий** в рамках *декомпозиции действия**, т. е построение плана решения задачи; указание области *действия*; указание условия инициирования *действия*; момент начала и завершения *действия*, в том числе планируемый и фактический, предполагаемая и/или фактическая длительность выполнения.

С зависимости от вида специфицируемого действия, решаемые системой задачи можно классифицировать на *информационные задачи* и *поведенческие задачи*.

С точки зрения формулировки поставленной задачи можно выделить *декларативные формулировки задачи* и *процедурные формулировки задачи*. Следует отметить, что данные классы задач не противопоставляются, и могут существовать формулировки задач, использующие оба подхода.

В рамках предлагаемого подхода вводятся и другие виды спецификации действий, которые подробнее рассмотрены в работе [5]

11 Уточнение понятия агента над общей семантической памятью

В рамках предлагаемого подхода единственным видом *субъектов*, выполняющих преобразования в *sc-памяти*, будем считать *sc-агенты*. Для формального определения понятия *sc-агента* воспользуемся введенным ранее понятием *класса логически атомарных действий*. Итак, будем называть *sc-агентом* некоторый *субъект*, способный выполнять *действия в sc-памяти*, принадлежащие некоторому определенному *классу логически атомарных действий*.

Логическая атомарность действий, выполняемых sc-агентом, предполагает, что каждый sc-агент реагирует на соответствующий ему класс ситуаций и/или событий, происходящих в sc-памяти, и осуществляет определенное преобразование sc-текста (текста SC-кода), находящегося в семантической окрестности обрабатываемой ситуации и/или события. При этом каждый sc-агент в общем случае не имеет информации о том, какие еще sc-агенты в данный момент присутствуют в системе, и осуществляет взаимодействие в другими sc-агентами исключительно посредством формирования каких-либо сообщений в общей sc-памяти. Таким сообщением может быть, например, вопрос, адресованный другим sc-агентам в системе (заранее не известно, каким конкретно) или ответ на вопрос, поставленный другими sc-агентами. Таким образом, каждый sc-агент в каждый момент времени

контролирует только фрагмент базы знаний в контексте решаемой данным агентом задачи, состояние всей остальной базы знаний в общем случае непредсказуемо для sc-агента.

Перечислим достоинства предлагаемого подхода к организации обработки знаний:

- поскольку обработка осуществляется агентами, которые обмениваются сообщениями только через общую память, добавление нового агента или исключение (деактивация) одного или нескольких существующих агентов, как правило, не приводит к изменениям в других агентах, поскольку агенты не обмениваются сообщениями напрямую;
- часто агенты работают параллельно и независимо друг от друга, выполняя разные действия в sc-памяти; таким образом, даже существенное расширение числа агентов в рамках одной системы не приводит к ухудшению ее производительности.

Поскольку предполагается, что копии одного и того же *sc-агента* (функционально эквивалентные *sc-агенты*) могут работать в разных системах, будучи при этом физически разными sc-агентами, то целесообразно рассматривать свойства и типологию не sc-агентов, а классов функционально эквивалентных sc-агентов, которые будем называть ***абстрактными sc-агентами***.

Каждый ***абстрактный sc-агент*** имеет соответствующую ему спецификацию. В спецификацию каждого ***абстрактного sc-агента*** входит: указание ключевых *sc-элементов* этого *sc-агента*; формальное описание условий инициирования данного *sc-агента*, т. е. тех *ситуаций* в *sc-памяти*, которые инициируют деятельность данного *sc-агента*; формальное описание первичного условия инициирования данного *sc-агента*, т. е. такой ситуации в *sc-памяти*, которая побуждает *sc-агента* перейти в активное состояние и начать проверку наличия своего полного условия инициирования; строгое, полное, однозначно понимаемое описание деятельности данного *sc-агента*, оформленное при помощи каких-либо понятных, общепринятых средств, не требующих специального изучения, например, на естественном языке; описание результатов выполнения работы соответствующих *sc-агентов.*

Более подробно понятие sc-агента рассмотрено в работе [5].

12 Уточнение понятия машины обработки базы знаний

В рамках предлагаемого подхода семантическая модель машины обработки базы знаний трактуется как *неатомарный абстрактный sc-агент*, являющийся результатом объединения всех *абстрактных sc-агентов*, входящих в состав какой-либо конкретной компьютерной системы, в один. Другими словами, под семантической моделью машины обработки базы знаний понимается коллектив всех sc-агентов, входящих в состав заданной компьютерной системы, воспринимаемый как единое целое.

Таким образом, можно выделить несколько основных уровней детализации любой машины обработки базы знаний: уровень самой машины обработки базы знаний; уровень неатомарных sc-агентов, входящих в состав машины, в том числе – более частных машин обработки базы знаний; уровень атомарных sc-агентов; уровень программ, реализующих алгоритмы деятельности соответствующих агентов.

Такая иерархия уровней позволяет говорить, во-первых, о возможности компонентного поэтапного создания машины обработки базы знаний, во-вторых – о возможности проектирования, отладки и верификации компонентов на разных уровнях независимо от других уровней, что существенно упрощает задачу создания машины обработки базы знаний за счет снижения накладных расходов.

Более подробно предлагаемая семантическая модель машины обработки базы знаний рассмотрена в [5].

13 Заключение

Семантические модели баз знаний и машин обработки баз знаний интеллектуальных систем, ориентированных на решение комплексных задач, являются не только объектами научных исследований, но и объектами проектирования. Это предполагает проведение серьезных научных исследований проектной деятельности, направленной на разработку интеллектуальных систем указанного класса.

Качество такой проектной деятельности определяется не только качеством разрабатываемых систем, но и минимально возможными сроками разработки, трудоемкостью разработки, требуемой квалификацией разработчиков, а также гибкостью (реконфигурируемостью) разрабатываемых систем.

Кроме того, для рассматриваемого класса интеллектуальных систем важна не только совместимость (интегрируемость) различных видов знаний и различных моделей решения задач в рамках одной системы, но и совместимость (интегрируемость) целых систем.

Каждая интеллектуальная система, ориентированная на решение комплексных задач и построенная на основе семантических моделей баз знаний и машин обработки баз знаний, имеет уникальную базу знаний и в общем случае уникальную машину обработки этой базы знаний, но и база знаний, и машина обработки этой базы знаний содержат большое количество многократно используемых компонентов.

Разрабатываемая нами Технология OSTIS (Open Semantic Technology for Intelligent Systems) [10], [3] как раз и направлена на решение указанных выше проблем.

Весь комплекс информационных и инструментальных средств поддержки проектирования интеллектуальных систем по Технологии OSTIS реализован в виде Метасистемы

IMS.ostis (Intelligent MetaSystem), которая сама также построена по Технологии OSTIS. Важным компонентом указанной метасистемы является библиотека многократно используемых компонентов проектируемых интеллектуальных систем.

Для создания технологии проектирования интеллектуальных систем, ориентированных на решение комплексных задач кроме обеспечения возможности совместного использования различных моделей представления и обработки знаний необходимо обеспечить гибкость (реконфигурируемость) баз знаний и машин обработки базы знаний и, как следствие, широкие возможности их постоянного совершенствования, а также создать библиотеки многократно используемых совместимых компонентов любого уровня сложности [5].

Литература

[1] Брюхов, Д.О., Ступников, С.А., Калиниченко Л.А. и др.: Извлечение информации из разноструктурированных данных и ее приведение к целевой схеме. Аналитика и управление данными в областях с интенсивным использованием данных: XVIII межд. конф. DAMDID / RCDL'2015 (Обнинск, Россия, 13–16 окт. 2015 года) / под ред. Л. А. Калиниченко, С. О. Старкова. Обнинск: НИЯУ МИФИ, сс. 81-90 (2015)

[2] Oberle, D.: Semantic Management of Middleware. Springer, 268 p. (2006)

[3] Голенков, В.В., Гулякина, Н.А.: Проект открытой семантической технологии компонентного проектирования интеллектуальных систем. Часть 1: Принципы создания. Онтология проектирования, (1), сс. 42-64 (2014)

[4] Гракова, Н.В., Давыденко, И.Т., Сергиенко, Е.С. и др.: Средства структуризации семантических моделей баз знаний. Открытые семантические технологии проектирования интеллектуальных систем (OSTIS-2016): материалы VI межд. науч.-техн. конф. / БГУИР; под ред. В.В. Голенкова. Минск: БГУИР, сс. 93-106 (2016)

[5] Shunkevich, D.: Ontology-based Design of Knowledge Processing Machines. Open Semantic Technologies for Intelligent Systems: материалы межд. науч.-техн. конф./ редкол.: В.В. Голенков (отв. ред.) и др.; Вып. 1 (Минск, 16–18 февраля 2017 г.). Минск: БГУИР, сс. 73-94 (2017)

[6] Добров, Б.В., Иванов, В.В., Лукашевич Н.В. и др.: Онтологии и тезаурусы. Учебно-методическое пособие. Казань: Изд-во Казан. ун-та, 190 с. (2006)

[7] Гаврилова, Т.А., Кудрявцев, Д.В., Муромцев, Д.И. Инженерия знаний. Модели и методы: Учебник. СПб.: Издательство «Лань», 348 с. (2016)

[8] Осипов, Г.С.: Методы искусственного интеллекта. 2-ое издание. М.: Физматлит, 296 с. (2015)

[9] Скворцов, Н.А., Калиниченко, Л.А., Ковалев, Д.Ю. Концептуальное моделирование предметных областей с интенсивным использованием данных. Аналитика и управление данными в областях с интенсивным использованием данных: XVIII межд. конф. DAMDID / RCDL'2016 (Ершово, Россия, 11–14 октября 2016 года) / ред. Л. А. Калиниченко, Я. Манолопулос, С. О. Кузнецова. М.: Торус Пресс, сс. 7-15 (2016)

[10] База знаний IMS // Метасистема IMS [Электронный ресурс] (2017). http://www.ims.ostis.net

Using metagraph approach for complex domains description

© Valeriy M. Chernenkiy, © Yuriy E. Gapanyuk, © Georgiy I. Revunkov,

© Yuriy T. Kaganov, © Yuriy S. Fedorenko, © Svetlana V. Minakova

Bauman Moscow State Technical University,
Moscow, Russia

chernen@bmstu.ru, gapyu@bmstu.ru, revunkov@bmstu.ru,

kaganov.y.t@bmstu.ru, fedyura11235@mail.ru, morgana_93@mail.ru

Abstract. This paper proposes an approach for complex domains description using complex network models with emergence. The advantages of metagraph approach are discussed. The formal definitions of the metagraph data model and metagraph agent model is given. The examples of data metagraph and metagraph rule agent are discussed. The metagraph and hypergraph models comparison is given. It is shown that the hypergraph model does not fully implement the emergence principle. The metagraph and hypernetwork models comparison is given. It is shown that the metagraph model is more flexible than hypernetwork model. Two examples of complex domains description using metagraph approach are discussed: neural network representation and modeling the polypeptide chain synthesis. The textual representation of metagraph model using predicate approach is given.

Keywords: metagraph, metavertex, metaedge, hypergraph, hypernetwork, neural network, polypeptide chain, lambda architecture.

1 Introduction

Currently, models based on complex networks are increasingly used in various fields of science from mathematics and computer science to biology and sociology. This is not surprising because the domains are becoming more and more complex.

Therefore, now it is important to offer not only a model that is capable of storing and processing Big Data but also a model that is capable of handling the complexity of data. That is why the development of a universal model for complex domains description is an actual task.

One of the varieties of such models is "complex networks with emergence". The emergent element means a whole that cannot be separated into its component parts.

As far as the authors know, currently there are two "complex networks with emergence" models: hypernetworks and metagraphs. The hypernetwork model is mature and it helps to understand many aspects of complex networks with an emergence.

But from the author's point of view, the metagraph model is more flexible and convenient for use in information systems.

This paper discusses the metagraph model and compares it with other complex graph models.

2 Complex networks models comparison

In this section, the metagraph model will be formally described and it will be compared with hypergraph and hypernetwork models.

2.1 Metagraph model formalization

A metagraph is a kind of complex network model, proposed by A. Basu and R. Blanning [1] and then adapted for information systems description by the authors [2]. According to [2]: $MG = \langle V, MV, E, ME \rangle$, where MG – metagraph; V – set of metagraph vertices; MV – set of metagraph metavertices; E – set of metagraph edges, ME – set of metagraph metaedges.

A metagraph vertex is described by the set of attributes: $v_i = \{atr_k\}, v_i \in V$, where v_i – metagraph vertex; atr_k – attribute.

A metagraph edge is described by the set of attributes, the source and destination vertices and edge direction flag: $e_i = \langle v_S, v_E, eo, \{atr_k\} \rangle, e_i \in E, eo = true|false$, where e_i – metagraph edge; v_S – source vertex (metavertex) of the edge; v_E – destination vertex (metavertex) of the edge; eo – edge direction flag (*eo=true* – directed edge, *eo=false* – undirected edge); atr_k – attribute.

The metagraph fragment: $MG_i = \{ev_j\}, ev_j \in (V \cup MV \cup E \cup ME)$, where MG_i – metagraph fragment; ev_j – an element that belongs to union of vertices, metavertices, edges and metaedges.

The metagraph metavertex: $mv_i = \langle \{atr_k\}, MG_j \rangle, mv_i \in MV$, where mv_i – metagraph metavertex belongs to set of metagraph metavertices *MV*; atr_k – attribute, MG_j – metagraph fragment.

Thus, a metavertex in addition to the attributes includes a fragment of the metagraph. The presence of private attributes and connections for a metavertex is distinguishing feature of a metagraph. It makes the definition of metagraph to be holonic – a metavertex may include a number of lower-level elements and in turn, may be included in a number of higher level elements.

From the general system theory point of view, a metavertex is a special case of the manifestation of the emergence principle, which means that the metavertex with its private attributes and connections becomes a whole that cannot be separated into its component parts. The example of metagraph is shown in figure 1.

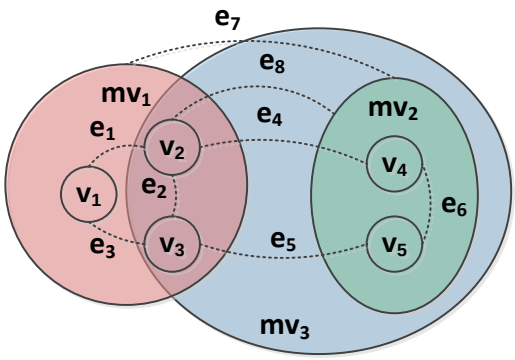


Figure 1 Example of metagraph

This example contains three metavertices: mv_1, mv_2, and mv_3. Metavertex mv_1 contains vertices v_1, v_2, v_3 and connecting them edges e_1, e_2, e_3. Metavertex mv_2 contains vertices v_4, v_5 and connecting them edge e_6. Edges e_4, e_5 are examples of edges connecting vertices v_2-v_4 and v_3-v_5 respectively and they are contained in different metavertices mv_1 and mv_2. Edge e_7 is an example of an edge connecting metavertices mv_1 and mv_2. Edge e_8 is an example of an edge connecting vertex v_2 and metavertex mv_2. Metavertex mv_3 contains metavertex mv_2, vertices v_2, v_3 and edge e_2 from metavertex mv_1 and also edges e_4, e_5, e_8 showing the holonic nature of the metagraph structure. Figure 1 shows that metagraph model allows describing complex data structures and it is the metavertex that allows implementing emergence principle in data structures.

The vertices, edges, and metavertices are used for data description and the metaedges are used for process description.

The metagraph metaedge: $me_i = \langle v_S, v_E, eo, \{atr_k\}, MG_j \rangle, me_i \in ME, eo = true|false$, where me_i – metagraph metaedge belongs to set of metagraph metaedges *ME*; v_S – source vertex (metavertex) of the metaedge; v_E – destination vertex (metavertex) of the metaedge; *eo* – metaedge direction flag (*eo=true* – directed metaedge, *eo=false* – undirected metaedge); atr_k – attribute, MG_j – metagraph fragment. The example of directed metaedge is shown in figure 2.

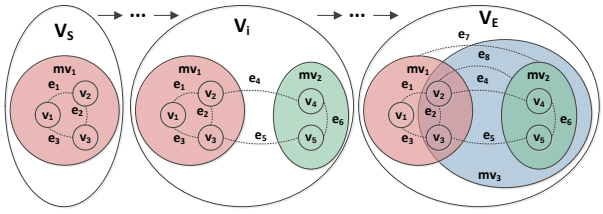


Figure 2 Example of directed metaedge

The directed metaedge contains metavertices v_S, … v_i, … v_E and connecting them edges. The source vertex contains a nested metagraph fragment. During the transition to the destination vertex, the nested metagraph fragment becomes more complex, as new vertices, edges, and metavertices are added. Thus, metaedge allows binding the stages of nested metagraph fragment development to the steps of the process described with metaedge.

2.2 Metagraph and hypergraph models comparison

In this section, the hypergraph model will be examined and compared with metagraph model. According to [3]: $HG = \langle V, HE \rangle, v_i \in V, he_j \in HE$, where HG – hypergraph; V – set of hypergraph vertices; HE – set of non-empty subsets of V called hyperedges; v_i – hypergraph vertex; he_j – hypergraph hyperedge.

A hypergraph may be directed or undirected. A hyperedge in an undirected hypergraph only includes vertices whereas, in a directed hypergraph, a hyperedge defines the order of traversal of vertices. The example of an undirected hypergraph is shown in figure 3.

This example contains three hyperedges: he_1, he_2, and he_3. Hyperedge he_1 contains vertices v_1, v_2, v_4, v_5. Hyperedge he_2 contains vertices v_2 and v_3. Hyperedge he_3 contains vertices v_4 and v_5. Hyperedges he_1 and he_2 have a common vertex v_2. All vertices of hyperedge he_3 are also vertices of hyperedge he_1.

Comparing metagraph and hypergraph models it should be noted that the metagraph model is more expressive then the hypergraph model. According to figures 1 and 3 it is possible to note some similarities between the metagraph metavertex and the hypergraph hyperedge, but the metagraph offers more details and clarity because the metavertex explicitly defines metavertices, vertices and edges inclusion, whereas the hyperedge does not. The inclusion of hyperedge he_3 in hyperedge he_1 in fig. 3 is only graphical and informal, because according to hypergraph definition a hyperedge inclusion operation is not explicitly defined.

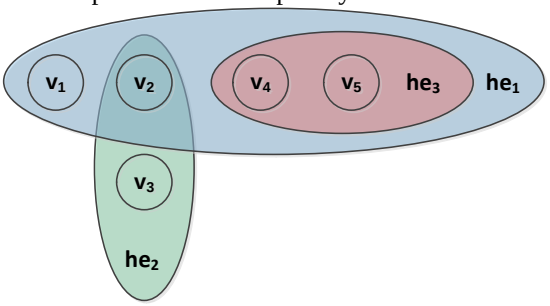


Figure 3 Example of undirected hypergraph

Thus the metagraph is a holonic graph model whereas the hypergraph is a near flat graph model that does not fully implement the emergence principle. Therefore, hypergraph model doesn't fit well for complex data structures description.

2.3 Metagraph and hypernetwork models comparison

The amazing fact is that the hypernetwork model was invented twice. The first time the hypernetwork model was invented by Professor Vladimir Popkov with colleagues in 1980s. Professor V. Popkov proposes several kinds of hypernetwork models with complex

formalization and therefore only main ideas of hypernetworks will be discussed in this section. According to [4] given the hypergraphs $PS \equiv WS_0, WS_1, WS_2, \cdots WS_K$. The hypergraph $PS \equiv WS_0$ is called primary network. The hypergraph WS_i is called a secondary network of order i. Also given the sequence of mappings between networks of different orders: $WS_K \xrightarrow{\Phi_K} WS_{K-1} \xrightarrow{\Phi_{K-1}} \cdots WS_1 \xrightarrow{\Phi_1} PS$. Then the hierarchical abstract hypernetwork of order K may be defined as $AS^K = \langle PS, WS_1, \cdots WS_K; \Phi_1, \cdots \Phi_K \rangle$. The emergence in this model occurs because of the mappings Φ_i between the layers of hypergraphs.

The second time the hypernetwork model was proposed by Professor Jeffrey Johnson in his monograph [5] in 2013. The main idea of Professor J. Johnson variant of hypernetwork model is the idea of hypersimplex (the term is adopted from polyhedral combinatorics). According to [5], a hypersimplex is an ordered set of vertices with an explicit n-ary relation and hypernetwork is a set of hypersimplices. In the hierarchical system, the hypersimplex combines k elements at the N level (base) with one element at the N+1 level (apex). Thus, hypersimplex establishes an emergence between two adjoining levels.

The example of hypernetwork that combines the ideas of two approaches is shown in figure 4.

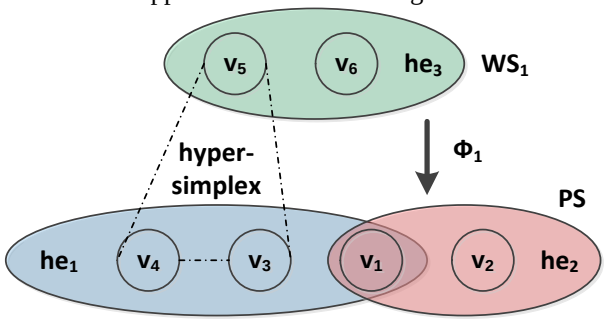


Figure 4 Example of hypernetwork

The primary network PS is formed by the vertices of hyperedges he_1 and he_2. The first level WS_1 of secondary network is formed by the vertices of hyperedge he_3. Mapping Φ_1 is shown with an arrow. The hypersimplex is emphasized with the dash-dotted line. The hypersimplex is formed by the base (vertices v_3 and v_4 of PS) and apex (vertex v_5 of WS_1).

The hypernetwork model became popular for complex domains description. For example, Professor Konstantin Anokhin [6] proposes a new fundamental theory of the organization of higher brain functions. According to this theory, biological neural networks (connectomes) are organized into cognitive hypernetworks (cognitomes). Vertices of cognitome form COGs (Gognitive Groups). Each COG may be represented as hypersimplex. The base of COG is a set of the vertices of underlying neural networks, and its apex is a vertex possessing a new quality at the macrolevel of cognitive hypernetworks. Thus, apex combines the base elements and emergence appears.

It should be noted that unlike the relatively simple hypergraph model the hypernetwork model is a full model with emergence. Consider the differences between the hypernetwork and metagraph models.

According to the definition of a hypernetwork it is a a layered description of graphs. It is assumed that the hypergraphs may be divided into homogeneous layers and then mapped with mappings or combined with hypersimplices. Metagraph approach is more flexible. It allows combining arbitrary elements that may be layered or not using metavertices.

Comparing the hypernetwork and metagraph models we can make the following notes:

- Hypernetwork model may be considered as "horizontal" or layer-oriented. The emergence appears between adjoining levels using hypersimplices. The metagraph model may be considered as "vertical" or aspect-oriented. The emergence appears between any levels using metavertices.
- In hypernetwork model, the elements are organized using hypergraphs inside layers and using mappings or hypersimplices between layers. In metagraph model, metavertices are used for organizing elements both inside layers and between layers. Hypersimplex may be considered as a special case of metavertex.
- Metagraph model allows organizing the results of previous organizations. The fragments of the flat graph may be organized into metavertices, metavertices may be organized in higher-level metavertices and so on. The metavertex organization is more flexible than hypersimplex organization because hypersimplex assumes base and apex usage and metavertex may include general form graph.
- Metavertex may represent a separate aspect of the organization. The same fragments of the flat graph may be included in different metavertices whether these metavertices are used for modeling different aspects.

Thus, we can draw a conclusion that metagraph model is more flexible than hypernetwork model.

However, it must be emphasized that the hypernetwork and metagraph models are only different formal descriptions of the same processes that occur in the networking with the emergence.

From the historical point of view, the hypernetwork model was the first complex network with an emergence model and it helps to understand many aspects of complex networks with an emergence.

3 Metagraph model processing

The metagraph model is designed for complex data and process description. But it is not intended for data transformation. To solve this issue, the metagraph agent (ag^{MG}) designed for data transformation is proposed. There are two kinds of metagraph agents: the metagraph function agent (ag^F) and the metagraph rule agent (ag^R). Thus $ag^{MG} = ag^F | ag^R$.

The metagraph function agent serves as a function with input and output parameter in form of metagraph:

$$ag^F = \langle MG_{IN}, MG_{OUT}, AST \rangle, \quad (1)$$

where ag^F – metagraph function agent; MG_{IN} – input parameter metagraph; MG_{OUT} – output parameter metagraph; AST – abstract syntax tree of metagraph function agent in form of metagraph.

The metagraph rule agent is rule-based: $\mathrm{ag}^{\mathrm{R}} = \langle \mathrm{MG}, \mathrm{R}, \mathrm{AG}^{\mathrm{ST}} \rangle, \mathrm{R} = \{\mathrm{r_i}\}, \mathrm{r_i}: \mathrm{MG_j} \to \mathrm{OP}^{\mathrm{MG}}$, where ag^R – metagraph rule agent; MG – working metagraph, a metagraph on the basis of which the rules of agent are performed; R – set of rules r_i; AG^{ST} – start condition (metagraph fragment for start rule check or start rule); MG_j – a metagraph fragment on the basis of which the rule is performed; OP^{MG} – set of actions performed on metagraph.

The antecedent of the rule is a condition over metagraph fragment, the consequent of the rule is a set of actions performed on metagraph. Rules can be divided into open and closed.

The consequent of the open rule is not permitted to change metagraph fragment occurring in rule antecedent. In this case, the input and output metagraph fragments may be separated. The open rule is similar to the template that generates the output metagraph based on the input metagraph.

The consequent of the closed rule is permitted to change metagraph fragment occurring in rule antecedent. The metagraph fragment changing in rule consequent cause to trigger the antecedents of other rules bound to the same metagraph fragment. But incorrectly designed closed rules system can lead to an infinite loop of metagraph rule agent.

If the agent contains only open rules it is called an open agent. If the agent contains only closed rules it is called a closed agent.

Thus, metagraph rule agent can generate the output metagraph based on the input metagraph (using open rules) or can modify the single metagraph (using closed rules). The example of metagraph rule agent is shown in figure 5.

The metagraph rule agent "metagraph rule agent 1" is represented as a metagraph metavertex. According to the definition it is bound to the working metagraph MG_1 – a metagraph on the basis of which the rules of the agent are performed. This binding is shown with edge e_4.

The metagraph rule agent description contains inner metavertices corresponds to agent rules (rule 1 … rule N). Each rule metavertex contains antecedent and consequent inner vertices. In given example mv_2 metavertex bound with antecedent which is shown with edge e_2 and mv_3 metavertex bound with consequent which is shown with edge e_3. Antecedent conditions and consequent actions are defined in form of attributes bound to antecedent and consequent corresponding vertices.

The start condition is given in form of attribute "start=true". If the start condition is defined as a start metagraph fragment then the edge bound start metagraph fragment to agent metavertex (edge e_1 in given example) is annotated with attribute "start=true". If the start condition is defined as a start rule than the rule metavertex is annotated with attribute "start=true" (rule 1 in given example). Figure 5 shows both cases corresponding to the start metagraph fragment and to the start rule.

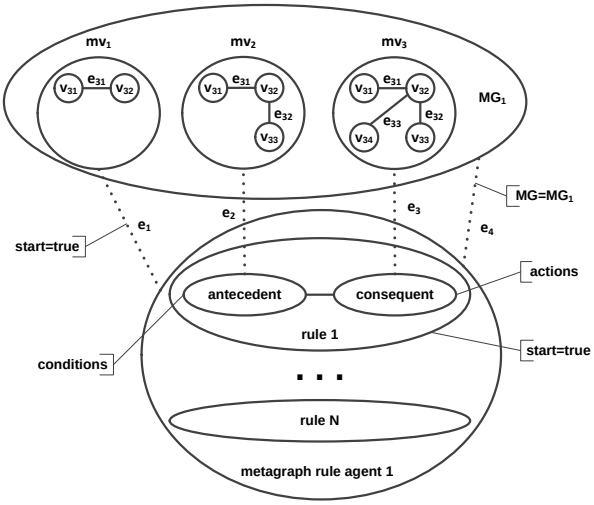


Figure 5 Example of metagraph rule agent

The distinguishing feature of metagraph agent is its homoiconicity which means that it can be a data structure for itself. This is due to the fact that according to definition metagraph agent may be represented as a set of metagraph fragments and this set can be combined in a single metagraph. Thus, the metagraph agent can change the structure of other metagraph agents.

4 The examples of complex domains description using metagraph approach

In this section, we give two examples of complex domains description using metagraph approach.

4.1 Using metagraph approach for neural network representation

This subsection is based on our paper [7]. We begin with simple perceptron representation using metagraph model. According to the Rosenblatt perceptron model [8], a conventional perceptron consists of three elements: S, A and R.

The layer of sensors (S) is an array of input signals. The associative layer (A) is a collection of intermediate elements which are triggered if a particular set of input signals is activated at the same time. The adder (R) is started when a particular collection of A-elements is activated concurrently.

According to the notation adopted in the M. Minsky and S. Papert perceptron model [8], the value of a signal on an A-element can be represented as a boolean predicate $\varphi(S)$, and the value of a signal in the adder layer as a predicate $\psi(A, W)$. According to [8], a function that takes either 0 or 1 is regarded as a boolean predicate.

Depending on the particular type of perceptron, the form of predicates $\varphi(S)$ and $\psi(A, W)$ can be different. Usually, predicate $\varphi(S)$ is used to check whether the total input signal from sensors exceeds a certain threshold or not. Also predicate $\psi(A, W)$ (where W is a weight vector) is used to see if the weighted sum from A-elements exceeds a particular threshold.

In our case, the actual form of predicates is not important. What is important is that the structure of *φ(S)* and *ψ(A, W)* can be represented as an abstract syntactic tree. Then we can represent the perceptron structure as a combination of metagraph function agents. Each predicate can be represented as a kind of the formula 1: $\varphi^F = \langle \mathrm{S}, \mathrm{A}, AST^{\varphi} \rangle, \psi^F = \langle \langle \{\varphi^F\}, W \rangle, \mathrm{R}, AST^{\psi} \rangle$. This representation is shown in figure 6.

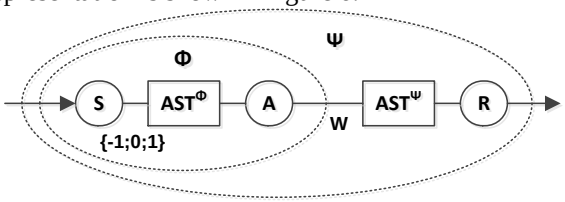


Figure 6 The perceptron representation as a combination of metagraph function agents

An A-element can be represented as a function agent φ^F. The input parameter is the value vector S, the output parameter is the value vector A. The description of the perceptron is similar to the description of the function agent ψ^F. The input parameter is a the metagraph representation of a tuple holding the description of A-elements as agent-functions φ^F and vector W. The output parameter is the amplitude of output signal R.

The description of functions can contain other parameters, e.g., threshold values, but we assume that these parameters are included in the description of the abstract syntactic tree.

Thus, we can describe the perceptron structure as a combination of metagraph function agents. Now we describe neural network operation using metagraph rule agents which are shown in figure 7.

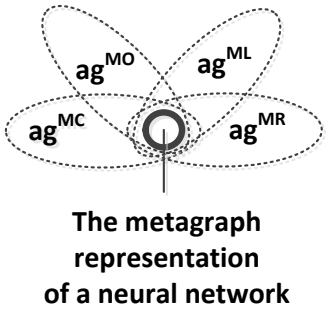


Figure 7 The structure of metagraph rule agents for neural network operation representation

The metagraph representation of neural network may be created similarly to the previously reviewed perceptron approach. Such a representation is a separate task that depends on neural network topology.

In order to provide a neural network operation the following agents are used:

- ag^{MC} – the agent responsible for the creation of the network;
- ag^{MO} – the agent responsible for the modification of the network;
- ag^{ML} – the agent responsible for the learning of the network;
- ag^{MR} – the agent responsible for the execution of the network.

In figure 7 the agents are shown as metavertices by dotted-line ovals.

The network-creating agent ag^{MC} implements the rules of creating an original neural network topology. The agent holds both the rules of creating separate neurons and rules of connecting neurons into a network. In particular, the agent generates abstract syntactic trees of metagraph function agents φ^F and ψ^F.

The network-modification agent ag^{MO} holds the rules of modification the network topology in process of operation. It is especially important for neural networks with variable topology such as HyperNEAT and SOINN.

The network-learning agent ag^{ML} implements a particular learning algorithm. As a result of learning the changed weights are written in the metagraph representation of the neural network. It is possible to implement a few learning algorithms by using different sets of rules for agent ag^{ML}.

The network-executing agent ag^{MR} is responsible for the start and operation of the trained neural network.

The agents can work separately or jointly which may be especially important in the case of variable topologies. For example when a HyperNEAT or SOINN network is trained, agent ag^{ML} can call the rules of agent ag^{MO} to change the network topology in the process of learning.

In fact, each agent uses its rules to implement a specific program "machine". The use of the metagraph approach allows us to implement the "multi-machine" principle: a few agents having different goals implement different operations on the same data structure.

Thus, we can draw a conclusion that metagraph approach helps to describe both the structure of separate neurons and the structure of neural network operation.

4.2 Using metagraph approach for modeling the polypeptide chain synthesis

Molecular biology is considered to be one of the most difficult to study topics of biological science. It's hard to believe that the complexity of functioning of the biological cell invisible to the human eye exceeds the complexity of functioning of a large ERP-system, which can contain thousands of business processes. The difficulty of studying biological processes is also due to the fact that in studying it is impossible to abstract from the physical and chemical features that accompany these processes. Therefore, the development of learning software that helps to better understand even one complex process is a valid task.

We will review the process of synthesis of a polypeptide chain which is also called "translation" in molecular biology. Translation is an essential part of the protein biosynthesis. This process is very valid from an educational point of view because protein biosynthesis is considered in almost all textbooks of molecular biology.

The translation process is very complicated and in this section, we review it in a simplified way.

The first main actor of the translation process is messenger RNA or mRNA, which may be represented as

a chain of codons. The second main actor of the translation process is ribosome consisting of the large subunit and the small subunit. The small subunit is responsible for reading information from mRNA and large subunit is responsible for generating fragments of the polypeptide chain.

According to [9] the translation process consists of three stages.

The first stage is initiation. At this stage, the ribosome assembles around the target mRNA. The small subunit is attached at the start codon.

The second stage is elongation. The small subunit reads information from the current codon. Using this information, the large subunit generates the fragment of the polypeptide chain. After that ribosome moves (translocates) to the next mRNA codon.

The third stage is termination. When the stop codon is reached, the ribosome releases the synthesized polypeptide chain. Under some conditions, the ribosome may be disassembled.

In this section, we use metagraph approach for translation process modeling. The representation is shown in figure 8.

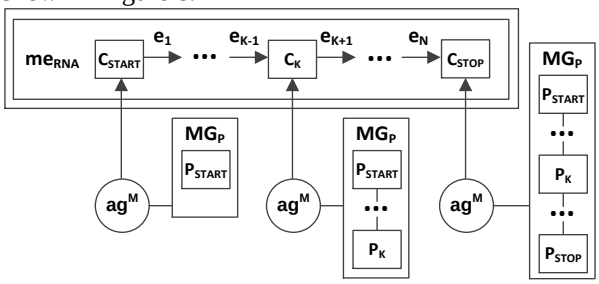


Figure 8 The representation of the polypeptide chain synthesis (translation) process based on metagraph approach

The mRNA is shown in figure 8 as metaedge $me_{RNA} = \langle C_{START}, C_{STOP}, eo = true, \{atr_k\}, MG_{RNA} \rangle$, where C_{START} – source metavertex (start codon); C_{STOP} – destination metavertex (stop codon); *eo=true* – directed metaedge; atr_k – attribute, MG_{RNA} – metagraph fragment, containing inner codons of mRNA (C_K) linked with edges.

The codon (shown in figure 8 as an elementary vertex) may also be represented as metavertex, containing inner vertices and edges according to the required level of detail.

Ribosome may be represented as metagraph rule agent $\mathrm{ag^{RB}} = \langle \mathrm{me_{RNA}}, \mathrm{R}, C_{START} \rangle, \mathrm{R} = \{\mathrm{r_i}\}, \mathrm{r_i}: \mathrm{C_K} \to \mathrm{P_K}$, where me_{RNA} – mRNA metaedge representation used as working metagraph; R – set of rules r_i; C_{START} – start codon used as start agent condition; C_K – codon on the basis of which the rule is performed; P_K – the added fragment of polypeptide chain.

The antecedent of the rule approximately corresponds to the small subunit of ribosome modeling. The consequent of the rule approximately corresponds to the large subunit of ribosome modeling.

Agent $\mathrm{ag^{RB}}$ is open agent generating output metagraph $\mathrm{MG_P}$ based on input metaedge me_{RNA}. The input and output metagraph fragments don't contain common elements.

While processing codons of mRNA agent $\mathrm{ag^{RB}}$ sequentially adds fragments of the polypeptide chain $\mathrm{P_K}$ to the output metagraph $\mathrm{MG_P}$. Vertices $\mathrm{P_K}$ are connected with undirected edges.

The process represented in figure 8 is very high-level. But metagraph approach allows representing related processes with different levels of abstraction.

For example, for each codon or peptide, we can link metavertex with its inner representation. And we can modify this representation during translation process using metagraph agents.

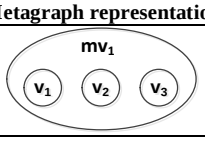
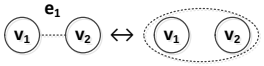
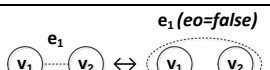
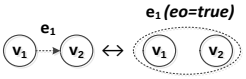
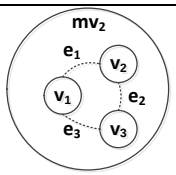
Thus, the metagraph approach allows us to represent a model of polypeptide chain synthesis which can be the basis for the learning software.

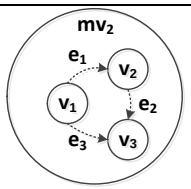
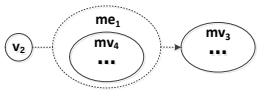
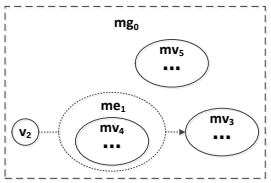
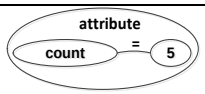
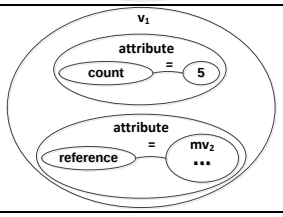
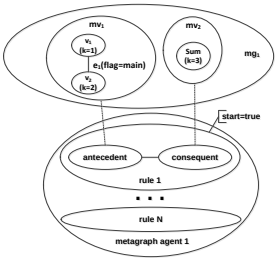
5 The textual representation of metagraph model

In previous sections, the formal definition and graphical examples of metagraph model were defined. But to successfully operate with metagraph model we also need textual representation. As such a representation, we use a logical predicate model that is close to logical programming languages e.g. Prolog. Logical predicates used in this section and boolean predicates used in subsection 4.1 should not be confused.

The classical Prolog uses following form of predicate: $predicate(atom_1, atom_2, \cdots, atom_N)$. We used an extended form of predicate where along with atoms predicate can also include key-value pairs and nested predicates: $predicate(atom, \cdots, key = value, \cdots, predicate(\cdots), \cdots)$. The mapping of metagraph model fragments into predicate representation is shown in Table 1.

Table 1 The textual representation of metagraph model

№	Metagraph representation	Textual representation
1		Metavertex(Name=mv_1, v_1, v_2, v_3)
2		Edge(Name=e_1, v_1, v_2)
3		Edge(Name=e_1, v_1, v_2, eo=false)
4		1. Edge(Name=e_1, v_1, v_2, eo=true) 2. Edge(Name=e_1, v_S=v_1, v_E=v_2, eo=true)
5		Metavertex(Name=mv_2, v_1, v_2, v_3, Edge (Name=e_1, v_1, v_2), Edge(Name=e_2, v_2, v_3), Edge(Name=e_3, v_1, v_3))

6		Metavertex(Name=mv_2, v_1, v_2, v_3, Edge(Name=e_1, v_S=v_1, v_E=v_2, eo=true), Edge(Name=e_2, v_S=v_2, v_E=v_3, eo=true), Edge(Name=e_3, v_S=v_1, v_E=v_3, eo=true))
7		Metaedge(Name=me_1, v_S=v_2, v_E=mv_3, Metavertex(Name=mv_4, …), eo=true)
8		Metagraph(Name=mg_0, Vertex(Name=v_2, …), Metavertex(Name=mv_3, …), Metavertex(Name=mv_5, …), Metaedge(Name=me_1, v_S=v_2, v_E=mv_3, Metavertex(Name=mv_4, …), eo=true))
9		Attribute(count, 5)
10		Vertex(Name=v_1, Attribute(count, 5), Attribute(reference, mv2))
11		Agent(Name= metagraph agent 1', WorkMetagraph=mg1, Rules(Rule(Name=rule 1', start=true, Condition (WorkMetagraph=mv_1, Vertex(Name=v_1, Attribute(k, $k1)), Vertex(Name=$v_2$, Attribute(k, $k2)), Edge($v_1$, v_2, Attribute(flag, main))) Action (WorkMetagraph=mv_2, Add(Vertex(Name=Sum, Attribute(k, Eval($k1+$k2)))))), Rule(…) …))

Case 1 shows the example of metavertex mv_1 which contains three nested disjoint vertices v_1, v_2, and v_3. The predicate corresponds to metavertex, nested vertices are isomorphic to atoms that are parameters of the predicate. As the name of the predicate, "Metavertex" is used as the corresponding element of metagraph model. Key-value parameter "Name" is used to set the name of metavertex. This case is the simplest, since nested vertices are disjoint, and metavertex in this case is isomorphic to the hypergraph hyperedge.

Case 2 shows metagraph edge which may be represented as a special case of metavertex containing source and destination vertices. This case is also isomorphic to the hypergraph hyperedge. The metagraph edge is represented as a predicate with the name "Edge". The source and destination vertices are represented as predicate atom parameters.

Case 3 also shows metagraph edge which fully complies with the formal definition of undirected edge including direction flag parameter.

Case 4 shows an example of directed edge. Direction flag parameter is also used. The source and destination vertices may be represented as predicate atom parameters (case 4.1) or as predicate key-value parameters (case 4.2).

Case 5 shows an example of metavertex mv_1 which contains three nested vertices v_1, v_2 and v_3 joined with undirected edges e_1, e_2, and e_3. Edges are represented with separate predicates that are nested to the metavertex predicate. Case 6 is similar to case 5 unless edges e_1, e_2, and e_3 are directed.

Case 7 shows an example of directed metaedge me_1 which joins vertex v_2 and metavertex mv_3 and contains metavertex mv_4. The metaedge is represented as a predicate with the name "Metaedge".

Case 8 shows an example of metagraph fragment mg_0 which contains vertex v_2, metavertices mv_3 and mv_5 and metaedge me_1 which joins vertex v_2 and metavertex mv_3 and contains metavertex mv4. The metagraph fragment is represented as a predicate with the name "Metagraph", the vertex as a predicate with the name "Vertex".

The attribute may be represented as a special case of metavertex containing name and value. Case 9 shows simple numeric attribute representation. Case 10 shows an example of vertex v_1 containing numeric attribute and reference attribute that refers to the metavertex mv_2. The attribute is represented as a predicate with the name "Attribute".

Case 11 shows an example of metagraph rule agent "metagraph agent 1" representation (the predicate with the name "Agent" is used). As a work metagraph mg_1 is used (parameter "WorkMetagraph"). The "Rules" predicate contains rules definition (nested predicate "Rule" is used). As a start rule "rule 1" is used which is defined by "start=true" parameter. Predicate "Condition" corresponds to the rule condition. Parameter "WorkMetagraph" contains a reference to the tested metavertex mv_1. The condition tests that metavertex mv_1 contains vertices v_1 and v_2 with attribute k. Founded values of k attribute of vertices v_1 and v_2 are assigned to the $k1 and $k2 variables. Vertices v_1 and v_2 should be joined with edge containing attribute "flag=main". If condition holds and metagraph fragment is found then actions are performed (actions are defined by predicate "Action"). Parameter "WorkMetagraph" contains a reference to the result metavertex mv_2. The example action contains adding new elements (that is defined by predicate "Add"). The vertex "Sum" is added containing attribute "k=$k1+$k2". Predicate "Eval" is used to define the calculated expression.

Thus, we defined a predicate description of all the main elements of metagraph data model.

The proposed predicate model is homoiconic. Since predicate approach is used both for metagraph data model definition and for metagraph agents definition then high-level metagraph agents may change the structure of low-level metagraph agents by modifying their predicate definition.

The textual representation of metagraph model may be used for storing metagraph model elements in relational or NoSQL databases.

It should be noted that metagraph model is well compatible with the Big Data approach. Nowadays the lambda architecture described in [10] is considered to be a classic approach.

The textual representation of metagraph model is the base for processing metagraph data on all layers of the lambda architecture. On the batch layer, the textual representation is used for storing in master dataset. On the serving layer, the textual representation helps to construct the batch views. On the speed layer, the textual representation helps to construct the real-time views. Batch and real-time views may be constructed using metagraph agents.

6 Conclusion

Nowadays complex network models have become popular for complex domains description.

The metagraph model is a kind of complex network model. The emergence in metagraph model is established using metavertices and metaedges.

The hypergraph model does not fully implement the emergence principle.

The hypernetwork model fully implements the emergence principle using hypersimplices. The metagraph model is more flexible than hypernetwork model.

For metagraph model processing, the metagraph function agents and the metagraph rule agents are used.

Two examples of complex domains description using metagraph approach are discussed: neural network representation and modeling the polypeptide chain synthesis. Metagraph approach helps to describe complex domains in a unified way.

The textual representation of metagraph model may be used for storing metagraph model elements in relational or NoSQL databases.

The metagraph model is well compatible with the Big Data approach, in particular with the lambda architecture.

References

[1] Basu A., Blanning R. Metagraphs and their applications. Springer, New York (2007)

[2] E. Samohvalov, G. Revunkov, Yu. Gapanyuk, “Metagraphs for describing semantics and pragmatics of information systems,” in Herald of Bauman Moscow State Technical University, vol. 1(100), pp.83-99, (2015)

[3] Vitaly I. Voloshin. Introduction to Graph and Hypergraph Theory. Nova Science Publishers, Inc., (2009)

[4] Akhmediyarova, A.T., Kuandykova, J.R., Kubekov, B.S., Utepbergenov, I.T., Popkov, V.K.: Objective of Modeling and Computation of City Electric Transportation Networks Properties. In: International Conference on Information Science and Management Engineering (Icisme 2015), pp. 106–111, Destech Publications, Phuket (2015)

[5] Johnson, J.: Hypernetworks in the Science of Complex Systems. Imperial College Press, London (2013)

[6] Anokhin, K.V.: Cognitom: theory of realized degrees of freedom in the brain. In: The Report Given at Fifth International Conference on Cognitive Science, Kaliningrad (2012)

[7] Fedorenko, Yu.S., Gapanyuk, Yu.E. Multilevel neural net adaptive models using the metagraph approach. Optical Memory and Neural Networks. Volume 25, Issue 4, pp. 228–235 (2016)

[8] Minsky, M.L., Papert, S.A. Perceptrons. The MIT Press (1988)

[9] Samish, I. Computational Protein Design. Springer Science+Business Media, New York (2017)

[10] Marz, N., Warren, J. Big Data. Principles and best practices of scalable realtime data systems. Manning, New York (2015)

Data Mining and Visualization: Meteorological Parameters and Gas Concentration Use Case

© Yas A. Alsultanny

Arabian Gulf University
Manama, Kingdom of Bahrain

alsultanny@hotmail.com

Abstract. Knowledge extraction from big data is one of the important subjects now and in future. Mining in the big data needs many steps, which must be implemented very carefully. The final step in big data mining is visualizing the results or summarizing the results numerically. This paper aims to mining the big data recorded by environmental station. These stations are recording the concentrations of some gases and meteorological parameters. The 2D and 3D data visualization is used to evaluate the capability of visualization in determining the effect of meteorological parameters on some gases that caused pollution. The results showing the visualization is a very important tool, and visualization can be used in mining big data, by showing the concentrations of gases. The paper recommends using big data visualization periodically as an alarming tool for monitoring the levels of pollution gases concentration.

Keywords: metrological parameters, gases concentration, filtering; preprocessing, decision tree, meteorological parameters.

1 Introduction

Big Data Mining (BDM) and Data Visualization (DV) are two important hot topics in the field of knowledge discovery. The big data can be visualized and analyzed to extract knowledge. The visual analytical tools have steadily improved during the last years in order to work with big data. The data collected from different resources, such as the station for monitoring pollution gases. These stations usually have an hourly readings to measure concentrations of gases such as; ozone O_3, nitrogen dioxide NO_2, sulfur dioxide SO_2, carbon monoxide CO, carbon dioxide CO_2, particulate matter (PM_{10} and $PM_{2.5}$), moreover these stations have hourly readings for meteorological parameters such as; Temperature (Temp), Humidity (Hu), Wind Speed (WS), Wind Direction (WD), and Air Pressure (AP).

Big data is a term used to describe some of current directions in information technology, as a concept that take into consideration data analysis. The amount of data in the world is huge, and it grows in an annual basis of 50% of its original size [1]. It is important to note that most of the big data is unstructured data, where it is not organized and does not fit the usual databases [2]. Big data can be used as a useful tool to enhance decision making [3].

Data Mining is the technique to get useful knowledge out of databases; data mining requires pre-processing and analytic approach for finding the value. Data mining requires many operations such as data integration, data selection, and so on [4].

Visual analytic first defined by Tomas and Cook in 2005 [5] as; the science of analytical reasoning facility by interactive visual interface. Murray in 2013 [6] described Data Visualization as; "fortunately, we humans are intensely visual creatures. Few of us can detect patterns among rows of numbers, but even young children can interpret bar charts, extracting meaning from those numbers' visual representations. Visualizing data is the fastest way to communicate it to others".

Air pollution is important in our life; most of the pollutants in the air are a result of emissions from cars, trucks, buses, factories, refineries, and other sources. The objective of this paper is to highlight the aspects of Big Data miming to visualize air pollution concentrations and it is relative to meteorological parameters.

2 Literature Review

Big data rises with the huge growth of data. It refers to the storing, processing, and analyzing the vast amounts of data. Big data brings new challenges to visualization because of the speed, size and diversity of data. One of the most common definitions of big data is data that have volume, variety, and velocity [7-9]. The term "Big Data" is surrounded by a lot of advertising, where many software vendors claim to have the ability to handle big data with their products [10]. Innovations in hardware technology such as those in network bandwidth, memory, and storage technology have assisted the technology of Big Data. The new innovations coupled with the latent need to analyze the massive unstructured data that stimulated their development [11].

Data Mining is the field of discovering novel and potentially useful information from large amounts of data [12]. Data mining defined as the use of analytical tools to discover knowledge in a database. The analytical tools may include machine learning, statistics, artificial intelligence, and information visualization [13]. Data mining categorized into seven categories as Fayyad et al. in 1996 [14] stated. These categories are regression, clustering, summarization, dependency modeling, link analysis, and sequence analysis. Knowledge Discovery

Proceedings of the XIX International Conference "Data Analytics and Management in Data Intensive Domains" (DAMDID/RCDL'2017), Moscow, Russia, October 10-13, 2017

in Databases (KDD) is the processing steps used to extract useful information from large collections of data [15]. Data mining manly has two methods: classification is assigns items in a collection to target categories or classes, and clustering is a form of unstructured learning method. Decision trees are types of classifications such as: Reduced Error Pruning (REP) tree, K Nearest Neighbors (KNN), the J48 based on C4.5 algorithm, and M5P algorithm is an improvement of the Quinlan's M5 algorithm [16-20].

"To visualize" has two meanings. "To form a mental image of something" refers to a cognitive, internal aspect whereas "to make something visible to the eye" refers to an external, perceptual role [21]. Visualization is any kind of technique to present information [22-23]. Data visualization refers to any graphic representation that can examine or communicate the data in any discipline [24]. The 3D visualization is gradually becoming the main trend in many fields including population gases and meteorological parameters [25].

3 Data Visualization

This study proposes a visualization method to represent graphically air pollution big data, to be an efficient method for knowledge discovery. This visual methodology is useful for people who are working in field of air pollution to have an efficient readability and accuracy of data analysis. Data visualization is the use of computer for visual representations of data. It aims at helping decision maker to detect effectively into big data. Data visualization is an efficient and intuitively accessible approach to identify patterns in large and diverse data sets.

Gases and metrological parameters visualizations can have two goals: Explanatory and Exploratory. Gases and metrological parameters data are usually recorded by automatic stations at regular time intervals. Metrological data is typically multivariate that often consists of many dimensions. Air pollution is a major concern in any city through the world. The visualization technique is used to aid visual analysis of the air pollution problem, followed by metrological data for knowledge discovery.

There are many steps must be taken in order to prepare data for visualization, these steps are shown in Figure 1. The steps are: stations sensors adjustment, data recording, data filtering, data preprocessing, normalization, aggregation, and visualization.

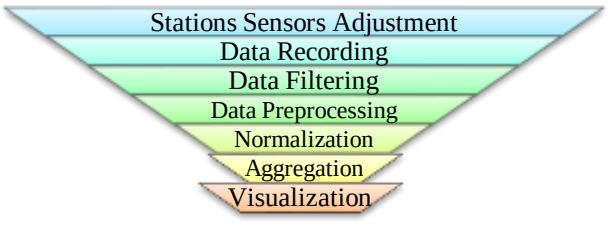


Figure 1 Big data acquisition and utilization

4 Data Collection and Analysis

The data available for this paper were collected from Arabian Gulf countries from one station in state of Kuwait; it was hourly time series data for eleven years, after data filtering and preprocessing, the data for one year 2015 was analyzed in this paper. The data represented on an hourly averaged reading, where the yearly readings for each gas or parameter must be 8,760 (24 hr*365 day), but the real readings after filtering and processing are 8,630, with 130 (1.5%) missed reading. The Rapidminer version 7.5 was used for processing and visualization the data of this paper.

Figure 2 shows the effect of temperature on the concentration of the five gases (O_3, NO_2, CO, CO_2, and SO_2) and PM_{10}. The figure visualizes the data distribution by using two-dimensional diagrams; the temperature has an opposite effect on O_3 and NO_2. The concentration of O_3 increased directly during the hottest hours, when the temperature was above 40ºC. While the temperature had a reverse effect on NO_2, the concentration of this gas became lower during the hottest hours, and its concentration was in its lightest levels, when the temperature was less than 10Cº. The effect of temperature on CO and CO_2 is very limited and this is clear from the figure, this indicates the temperature has no effect on these two gases. The hottest hours have a direct effect on SO_2 and PM_{10}, their concentrations usually increased during summer and especially in the hottest hours of a day.

Figure 3 shows the effect of humidity on the five gases and PM_{10}. The humidity has a reverse effect on O_3 and NO_2, their concentrations are increased with lower concentration of humidity, moreover the concentrations of CO, CO_2, and SO_2 increased with lower percentage of humidity. The PM_{10} concentration significantly reduced, when the humidity percentage was higher than 70%. These results are true, because the highest percentages of humidity, reducing the five gases and PM_{10} disperse.

Figure 4 shows the three dimensions scatter diagrams to visualize the effect of both temperature and humidity at the same time on the five gases and PM10. The figure shows again most of the readings of O3 are concentrated in the region of hottest temperature and low percentage of humidity. The concentrations of NO2 increased at the lowest temperature and humidity. For CO, CO_2, SO_2, and PM_{10} their readings are concentrated in the region of hottest temperature and low percentage of humidity.

A decision tree is a predictive model [26]. It was implemented in this paper to predicate PM_{10}, which is measured in part per million (ppm), by stating the effect of temperature and wind speed. To implement the decision tree the PM_{10}, temperature, and humidity were classified into: 0=0-50, 1=51-150, 2=151-400, 3=401-700, 4=701-1000, 5=1001-1500, 6=1501-2500, 7=2501 and more. The temperature in centigram degree (Cº) classified into: 0=0-6, 1=7-11, 2=12-16, 3=17-21, 4=22-26, 5=27-35,6=36-46, 7=47 and more. The wind speed meter per second (m/s) classified into: 0=0-2, 1=3-5, 2=6-8, 3=9-12, 4=13 and more. The decision rules of the decision tree to predicate PM_{10}, as an example by using temperature and wind speed-readings are as follows.

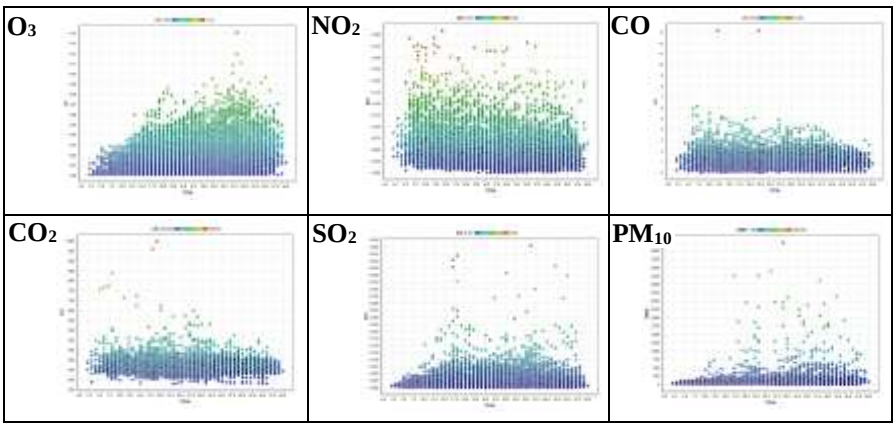


Figure 2 Effect of temperature on the five gases and PM_{10}

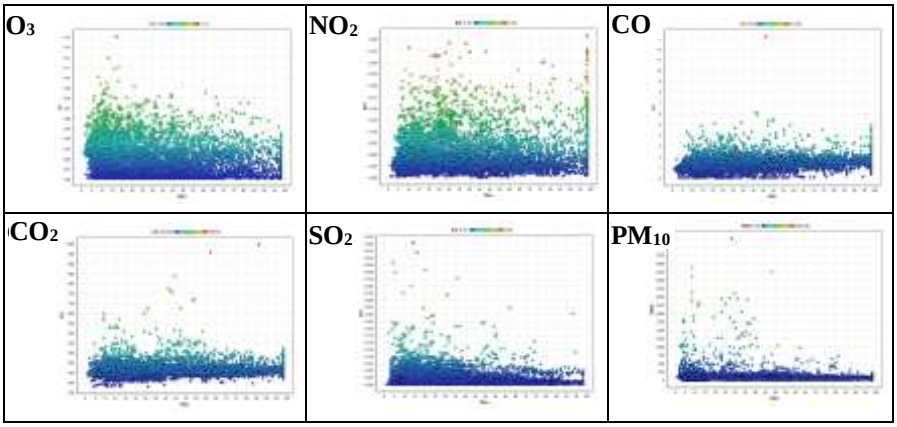


Figure 3 Effect of humidity on the five gases and PM_{10}

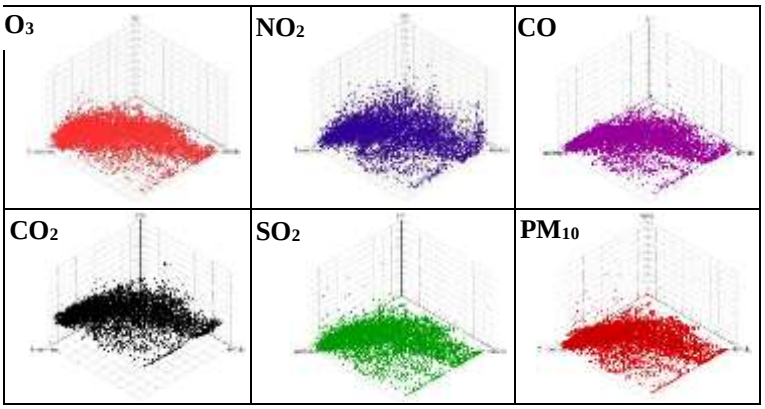


Figure 4 Effect of temperature and humidity on the five gases and PM_{10}

It shows when wind speed between 6-12m/s and temperature 22=35 C°, the PM_{10} will be between 151-400 ppm.

Tree

WS > 3.500: 2 {1=0, 0=0, 2=2, 3=2, 7=0, 4=0, 5=1, 6=0}
WS ≤ 3.500
| WS > 2.500
| | Temp > 4.500: 2 {1=15, 0=0, 2=28, 3=8, 7=0, 4=4, 5=4, 6=5}
| | Temp ≤ 4.500: 1 {1=41, 0=11, 2=16, 3=0, 7=2, 4=0, 5=2, 6=3}
| WS ≤ 2.500: 1 {1=4802, 0=1472, 2=634, 3=77, 7=3, 4=23, 5=16, 6=10}

5 Conclusion

The problems of storing and analysis of big data are facing all the organization through the world, especially the environmental organizations taking interest in monitoring pollution gases. These organizations have one or more online reading stations installed near industrial cities and oil refinery stations.

Using the 2D and 3D scatter diagram to visualize the data reading is one of the important tools. That can be used by decision makers to explore the concentration

of pollutant gases and effect of meteorological parameters, by using these types of visualization the decision makers can take their decision in stopping or reducing the working hours of the factories or refinery stations that cause the major pollution.

We recommend each factory of refinery, using the same methods of visualizing the pollution gases to take their decision to stop their factory of refinery station or reducing the hours of working hours, when the temperature rises to more that 45C°.

References

[1] Gantz, J., Reinsel, D.: Extracting value from chaos. IDC IVIEW. https://www.emc.com/collateral/analyst-reports/idc-extracting-value-from-chaos-ar.pdf

[2] Lohr, S.: The age of big data. http://www.nytimes.com/2012/02/12/sunday-review/big-datas-impact-in-the-world.html

[3] Shumway R.: One solution for air pollution: big data. http://www.deseretnews.com/article/865617771/One-solution-for-air-pollution-Big-data.html

[4] Han, J., Kamber, M., Jian, P.: Data mining: concepts and techniques. Elsevier Inc., (2012)

[5] Thomas, J., Cook, J.: Illuminating the path: the research and development agenda for visual analytics. National Visualization and Analytics Center (2005)

[6] Murray, S.: Interactive data visualization for the web. O'Reilly Media, Inc. (2013)

[7] http://www.sas.com/en_us/home.html

[8] De Mauro, A., Greco, M., Grimaldi, M.: Grimaldi formal definition of big data based on its essential features. Journal of Library Review, vol. 65, no. 3, pp. 122–135 (2016)

[9] Dion, M., AbdelMalik, P., Mawudeku, A.: Big data and the Global Public Health Intelligence Network (GPHIN). vol. 41, pp. 209-219 (2015)

[10] Heudecker, N., Beyer, A., Laney, D., Cantara, M., White, A., Edjlali, R., McIntyre, A.: Predicts 2014: big data. gartner insight. Gartner Research, Stanford, Connecticut (2013)

[11] Bhagattjee, B.: Emergence and taxonomy of big data as a service. Working Paper CISL# 2014-06. Massachusetts Institute of Technology (2014)

[12] Cheng, S., Liu, Shi, Y., Jin, Y., Li, B.: Evolutionary computation and big data: key challenges and future directions. Proceedings of the First International Conference on Data Mining and Big Data, Bali. Indonesia, pp 3-14, June 25-30 (2016)

[13] Redpath, R.: A comparative study of visualization techniques for data mining. MSc thesis. School of Computer Science and Software Engineering, Monash University, Australia (2000)

[14] Fayyad, U., Piatetsky-Shapiro, G., Smyth, P.: From data mining to knowledge discovery in databases. American Association for Artificial Intelligence, pp. 37-54 (1996)

[15] Frawley, J. Piatetsky-Shapiro, G., Matheus, J.: Knowledge discovery in databases: an overview; knowledge discovery in databases. AAAI Press/ The MIT Press, Menlo Park, California, USA (1991)

[16] Tan, P., Steinbach, M., Kumar, V.: Introduction to data mining. Pearson Addison Wesley (2006)

[17] Witten, I., Frank, E., Hall, M., Pal, C.:Data mining: practical machine learning tools and techniques. Elsevier Inc., 4th Edition (2017)

[18] Kantardzic, M.: Data mining: concepts, models, methods, and algorithms. Johm Wiley and Sons Inc., 2nd Edition (2011)

[19] Masethe, M., Masethe, H.: Prediction of work integrated learning placement using data mining algorithms. Proceedings of the World Congress on Engineering and Computer Science, San Francisco, USA, vol I, WCECS 2014, 22-24 October (2014)

[20] Neeb, H., Kurrus, C.: Distributed K-nearest neighbors. https://stanford.edu/~rezab/classes/cme323/S16/projects_reports/neeb_kurrus.pdf

[21] Oxford English Dictionary, Visualization. Oxford University Press (2009)

[22] Chen, C., Hardle, W., Unwin, A.: Handbook of data visualization. Springer (2008)

[23] Keim, A., Mansmann, J., Thomas, S., Ziegler, H.: Visual analytics: scope and challenges. Berlin, Heidelberg, Springer-Verlag (2008)

[24] Few, S.: Now you see it: simple visualization techniques for quantitative analysis. Analytics Press, Oakland (2009)

[25] NESSI.: Big data a new world of opportunities. White Paper (2012)

[26] Rokach, L., Maimon, O.: Data mining with decision trees: theory and applications. World Scientific Publishing (2008)

Подходы к решению задач в ОИИД

Approaches for problem solving in DID

Верификация процесса конвективной диффузии на основе анализа многомерных временных рядов

© М.Г. Матвеев © Е.А. Сирота © М.Е. Семенов © А.В. Копытин

Воронежский государственный университет,
Воронеж, Россия

mgmatveev@yandex.ru atoris@list.ru

Аннотация. Предложен метод верификации процесса конвективной диффузии на основе анализа многомерных временных рядов. Он основан на сравнительном анализе конечно-разностных представлений исследуемой модели и авторегрессионного описания временных рядов наблюдений. Метод включает получение МНК-оценок параметров многомерной авторегрессии и построение вариантов систем алгебраических уравнений, связывающих оценки авторегрессии и параметры соответствующих дифференциальных уравнений. Система алгебраических уравнений, которой удовлетворяют полученные оценки, определяет структуру модели и соответствующие значения параметров дифференциального уравнения. Приведен численный пример идентификации процесса изменения температуры атмосферного воздуха.

Ключевые слова: уравнения в частных производных; структурная идентификация; параметрическая идентификация; многомерная авторегрессия; статистические гипотезы.

Verification of the Convective Diffusion Process Based on the Analysis of Multidimensional Time Series

© M.G. Matveev © E.A. Sirota © M.E. Semenov © A.V. Kopytin

Voronezh State University,
Voronezh, Russia

mgmatveev@yandex.ru atoris@list.ru

Abstract. A method for verifying the process of convective diffusion based on the analysis of multidimensional time series is proposed. The proposed method is based on a comparative analysis of finite-difference representations of the model under study and an autoregressive description of time series of observations. The method includes the LSM (Least square method) estimates of the parameters of multidimensional autoregression and the construction of versions of systems of algebraic equations connecting the estimates of autoregression and the parameters of the corresponding differential equations. The system of algebraic equations satisfied by the obtained estimates determines the structure of the model and the corresponding values of the parameters of differential equation. A numerical example of identifying the process of changing the temperature of atmospheric air is given.

Keywords: partial differential equations, structural identification, parametric identification, multidimensional autoregression, statistical hypothesis.

1 Введение

Математические модели физических процессов основаны на априорном предположении о механизме функционирования этих процессов [1]. Однако такие предположения не всегда в достаточной степени обоснованы. Ошибки спецификации модели возникают как на структурном, так и на параметрическом уровнях. Эффективным инструментом проверки предполагаемых механизмов процессов могут служить методы анализа временных рядов, образующихся при наблюдении за параметрами функционирования этих процессов [2, 3]. Предлагаемый метод основан на сравнительном анализе конечно-разностных представлений исследуемой модели процесса и авторегрессионнного описания временных рядов наблюдений.

Будем рассматривать широкий класс пространственно-распределенных динамических систем, для которых характерны диффузионные процессы, процессы адвекции или их сочетание. Соответствующее дифференциальное уравнение в

Труды XIX Международной конференции «Аналитика и управление данными в областях с интенсивным использованием данных» (DAMDID/ RCDL'2017), Москва, Россия, 10–13 октября 2017 года

частных производных с начальными и граничными условиями имеет следующий общий вид:

$$\frac{\partial y}{\partial t} + \vartheta \frac{\partial y}{\partial l} = D \frac{\partial^2 y}{\partial l^2}, \quad (1)$$

$$y(0, l) = \varphi(l), y(t, l^{\min}) = f_1(t), y(t, l^{\max}) = f_1(t),$$

где ϑ – скорость адвекции, D – коэффициент диффузии, l – пространственная координата.

Источником информации о поведении системы являются данные натурных измерений переменной y_i^t с погрешностью ε_i^t в виде «белого шума» – $x_i^t = y_i^t + \varepsilon_i^t$ в последовательные моменты времени $t = 0,1,\dots$ в узлах одномерной пространственной регулярной сетки $i = 0,1,\dots,n,$ т. е. многомерный временной ряд. Рассмотрение одномерной сетки ничем не ограничивает дальнейшие исследования, зато позволяет избежать громоздких построений, характерных для плоских и объемных пространств.

Пусть в рассматриваемой системе могут протекать процессы диффузии и адвекции. Утверждать, что поведение системы определяется одним из указанных процессов или действуют оба процесса одновременно, нет достаточных оснований. Также неизвестны параметры (1), которые рассматриваются как константы, полученные в результате усреднения по времени.

Задача заключается в верификации процессов конвективной диффузии на основе анализа многомерных временных рядов и разработке алгоритмов структурной и параметрической идентификации механистической модели с постоянными коэффициентами по наблюдаемым значениям x_i^t.

2 Разностные схемы для вариантов уравнений механистической модели

Для решения задачи составим явные трехточечные разностные схемы для уравнений каждого из вариантов структуры процессов:

диффузия и адвекция

$$\frac{y_i^{t+1} - y_i^t}{\Delta t} + \vartheta \frac{y_{i+1}^t - y_{i-1}^t}{2\Delta l} = D \frac{y_{i+1}^t - 2y_i^t + y_{i-1}^t}{\Delta l^2}, \quad (2)$$

$$y_i^{t+1} = (b_1 + b_2) y_{i-1}^t + (1 - 2b_2) y_i^t + (b_2 - b_1) y_{i+1}^t;$$

$$b_1 = \frac{\vartheta \Delta t}{2\Delta l}; \quad b_2 = \frac{D\Delta t}{\Delta l^2};$$

диффузия в неподвижной среде

$$\frac{y_i^{t+1} - y_i^t}{\Delta t} = D \frac{y_{i+1}^t - 2y_i^t + y_{i-1}^t}{\Delta l^2};$$

$$y_i^{t+1} = b_2 y_{i-1}^t + (1 - 2b_2) y_i^t + b_2 y_{i+1}^t; \quad (3)$$

адвекция

$$\frac{y_i^{t+1} - y_i^t}{\Delta t} + \vartheta \frac{y_{i+1}^t - y_{i-1}^t}{2\Delta l} = 0, \quad (4)$$

$$y_i^{t+1} = b_1 y_{i-1}^t + y_i^t - b_1 y_{i+1}^t.$$

Представим разностные схемы (2)–(4) в виде обобщенной рекуррентной зависимости для произвольного узла i с начальными и граничными условиями:

$$y_i^{t+1} = \alpha_1 y_{i-1}^t + \alpha_2 y_i^t + \alpha_3 y_{i+1}^t \quad (5)$$

$$i = 2,\dots n-2; t = 0,1,\dots,$$

$$y^0 = (y_{i-1}^0; y_i^0; y_{i+1}^0),$$

$$y_{i-1} = \left(y_{i-1}^0; y_{i-1}^1; \dots; y_{i-1}^k\right),$$

$$y_{i+1} = \left(y_{i+1}^0; y_{i+1}^1; \dots; y_{i+1}^k\right),$$

где α_i – коэффициенты, вид которых определяется вариантом структуры процессов.

Начальные и граничные условия задаются как результаты натурных измерений, поэтому

$$y^0 = x^0 = (x_{i-1}^0; x_i^0; x_{i+1}^0),$$

$$y_{i-1} = x_{i-1} = \left(x_{i-1}^0; x_{i-1}^1; \dots; x_{i-1}^k\right),$$

$$y_{i+1} = x_{i+1} = \left(x_{i+1}^0; x_{i+1}^1; \dots; x_{i+1}^k\right).$$

Заметим, что в соответствии со свойством консервативности [4] сумма коэффициентов правой части выражений (2)–(4) равна единице. Соответственно $\sum_{j=1}^{3} \alpha_j = 1$.

3 Построение модели многомерного временного ряда и ее сравнение с разностной схемой

Моделирование многомерного временного ряда будем осуществлять в классе линейных стохастических моделей авторегрессии [3]. Допустим, что в каждом узле регулярной сетки протекает марковский процесс без последействия, и временные ряды в смежных узлах имеют высокие значения коэффициентов линейной корреляции, что обуславливает рассмотрение многомерных рядов. Такие допущения позволяют специфицировать стохастическую модель в i-м узле в виде

$$\tilde{x}_i^{t+1} = M\left(\frac{x_i^{t+1}}{x_{i-1}^t}, x_i^t, x_{i+1}^t\right) = \alpha_1 x_{i-1}^t + \alpha_2 x_i^t + \alpha_3 x_{i+1}^t;$$

$$i = 2,\dots n-2; \quad t = 0,1,\dots, \quad (6)$$

где α_1, α_2, α_3 – оценки параметров авторегрессии.

Вычисления по выражениям (5) и (6) существенно различаются, несмотря на их кажущееся сходство. Выражение (5) представляет собой рекуррентную формулу для вычисления модельных значений переменной y в трех попарно смежных узлах сетки, с условной устойчивостью, следовательно, и сходимостью к решению соответствующего дифференциального уравнения. Выражение (6) позволяет вычислять модельные значения случайной переменной x в тех же узлах, но автономно для каждого узла в каждый момент времени, как это определяется формулой (6). Различие в вычислительных алгоритмах (5) и (6) затрудняет сравнение результатов вычислений. В частности, для целей проводимого исследования важно показать, что математическое ожидание оценок a в авторегрессии (6) равно параметрам α в разностной схеме (5).

Найдем математические ожидания левой и правой части авторегрессионной зависимости, учитывая, что

математическое ожидание условного математического ожидания случайной величины есть математическое ожидание этой случайной величины:

$$M(x_i^{t+1}) = M(a_1)M(x_{i-1}^t) + M(a_2)M(x_i^t) + \\ + M(a_3)M(x_{i+1}^t).$$

Полученное выражение можно переписать с учетом граничных условий и очевидного условия

$$M(x_i^{t+1}) = M(y_i^{t+1}) + M(\varepsilon_i^{t+1}) = y_i^{t+1}$$

в виде

$$y_i^{t+1} = M(a_1)y_{i-1}^t + M(a_2)y_i^t + M(a_3)y_{i+1}^t,$$

что доказывает равенство математических ожиданий оценок параметров авторегрессии и параметров α разностной схемы (5). Полученный результат позволяет приравнять оценки параметров авторегрессии и параметров α и из полученных уравнений найти параметры b_1 и b_2.

Трем вариантом сочетания процессов будут соответствовать три варианта систем алгебраических уравнений относительно параметров b_1 и b_2 разностных схем (2)–(4):

для диффузии и адвекции

$$\begin{cases} b_1 + b_2 = a_1, \\ 1 - 2b_2 = a_2, \\ b_2 - b_1 = a_3; \end{cases} \quad (7)$$

для диффузии

$$\begin{cases} b_2 = a_1, \\ 1 - 2b_2 = a_2, \\ b_2 = a_3; \end{cases} \quad (8)$$

для адвекции

$$\begin{cases} b_1 = a_1, \\ 1 = a_2, \\ -b_1 = a_3. \end{cases} \quad (9)$$

В полученных уравнениях в правой части всюду стоят оценки, т. е. значения случайных величин. Следовательно, решение каждого из уравнений следует проводить как проверку соответствующей статистической гипотезы в предположении нормального распределения оценок. Вариант системы уравнений, согласующийся с соответствующей гипотезой, определит адекватный процесс. Решение с проверкой статистической гипотезы показано в примере, приведенном в следующем разделе.

4 Пример реализации предложенной методики

Диффузия и адвекция являются характерными процессами, определяющими изменение состояния земной атмосферы. В частности, именно эти процессы определяют динамику температурных полей на заданных изобарических поверхностях [8, 9]. Актуальной задачей метеорологии является определение структуры и параметров атмосферных процессов. Покажем, что решение этой задачи возможно, по крайней мере, для среднегодовых значений, по результатам регулярных наблюдений за температурой атмосферного воздуха над заданным участком земной поверхности.

Для экспериментальной апробации использовались статистические данные реанализа параметров атмосферы за 2012 год [10]. Эти данные представляют собой ежедневные значения температуры в узлах плоской регулярной сетки с шагом 2,5°. Рассматривались результаты наблюдений температуры при геопотенциале 300 ГПа в узле сетки с координатами 35° северной широты; 7,5° восточной долготы. По указанным данным была построена одномерная модель множественной авторегрессии

$$\tilde{x}_{7,5}^{t+1} = 1{,}03x_5^t - 1{,}14x_{7,5}^t + 1{,}11x_{10}^t x_{10}^t, \quad (10)$$

где $t = 0, 1, \ldots$. Заметим, что сумма коэффициентов в выражении (10) равна единице, что можно рассматривать как признак того, что временной ряд адекватно отражает динамику температуры, описанную разностной схемой (5). Для оценки процессов, оказывающих доминирующее влияние на динамику температуры, проведем статистический анализ систем (7)–(9).

Стандартные ошибки оценок равны $\sigma_{a_1} = 0{,}14$; $\sigma_{a_2} = 0{,}25$; $\sigma_{a_3} = 0{,}13$. Подставим полученные оценки a в систему (7), что соответствует принятию гипотезы о совместном влиянии на температурную динамику процессов диффузии и адвекции. Допустим, что ошибки оценок a равномерно распределены между коэффициентами b. Тогда последние два уравнения системы (7) позволяют вычислить оценки $b_2 = 1{,}07$ со стандартной ошибкой $\sigma_{b_2} = 0{,}07$ и $b_1 = -0{,}04$ со стандартной ошибкой $\sigma_{b_1} = 0{,}125$. Последнее равенство означает, что b_1 можно считать равным нулю. Поскольку коэффициент b_1 отвечает за адвекцию (см. раздел 2), нулевая гипотеза о совместном влиянии процессов диффузии и адвекции отклоняется.

Подставим полученные оценки в систему (8), что соответствует принятию нулевой гипотезы о влиянии диффузии. В этом случае нулевая гипотеза может быть принята, если подтвердятся гипотезы о равенстве a_1 и a_2 и тождественности второго равенства системы (8). Рассмотрим гипотезу о равенстве нулю разности a_1 и a_2. Эта разность составляет 0,08 при стандартной ошибке a_1 и $\sigma_{a2-a1} = 0.13 + 0.14 = 0.27$, что дает основание принять гипотезу о равенстве и улучшить оценку путем вычисления среднего значения $b_2 = (1.03 + 1.11)/2 = 1.07$. Для проверки гипотезы о тождественности второго равенства системы (8) подставим в него полученные оценки a_2 и b_2 и получим –1,14=–1,14, что при стандартной ошибке $\sigma_{a2} = 0.25$ позволяет принять гипотезу о тождественности, следовательно, и нулевую гипотезу о влиянии на динамику температуры процесса диффузии.

Подстановка оценок a в систему (9) для процесса адвекции позволяет сразу отклонить гипотезу о

влиянии на динамику температуры процесса адвекции из-за наличия противоречивых знаков оценок.

Окончательный вывод: на динамику температуры атмосферы в рассматриваемой точке в среднем в течение года доминирующее влияние оказывали процессы атмосферной диффузии.

Чтобы убедиться в адекватности результатов структурной и параметрической идентификации, было найдено численное решение уравнения диффузии по неявной разностной схеме, поскольку в данном случае условие устойчивости явной схемы не выполняется. Уровни наблюдаемого временного ряда температур, результаты моделирования с использованием авторегрессии (10) и решение уравнения диффузии с использованием неявной разностной схемы с найденным значением параметра $b_2 = 2.02$ представлены на рис. 1.

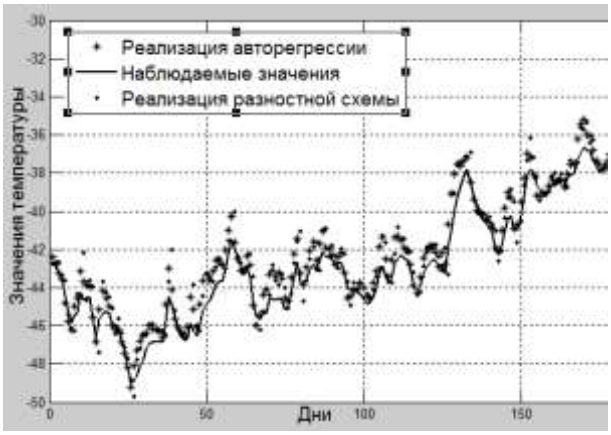


Рисунок 1 Временной ряд и результаты моделирования

Экспериментальным подтверждением правильности предложенной методики должно быть хорошее совпадение наблюдаемых значений температуры, значений, полученных по авторегрессионной модели (10), и численного решения уравнения диффузии. Качество авторегрессионной модели и модели диффузии характеризуется коэффициентом детерминации R^2. В обоих случаях этот коэффициент равен 0,78, что говорит об одинаковых качественных характеристиках моделей. Степень близости решений, полученных по этим моделям можно оценить коэффициентом парной корреляции, который составил в нашем случае 0,97, что еще раз подтверждает практическую эквивалентность найденных решений.

5 Заключение

Для идентификации динамических моделей с распределенными переменными обычно используются либо непараметрические методы идентификации, например, на основе результатов активного эксперимента [11], который не всегда возможен, либо параметрические методы с аппроксимацией производных [12]. В последнем случае возникает опасность неадекватного приближения производных в условиях помех [13]. Различные статистические методы были разработаны для оценки параметров моделей, описываемых обыкновенными дифференциальными уравнениями. Так, например, в [14–16] предложены иерархические байесовские подходы к этой проблеме. Эти методы требуют неоднократного численного решения соответствующих обыкновенных дифференциальных уравнений, что в ряде случаев требует значительного временного ресурса. Для оценки постоянных параметров модели в [17] предложен обобщенный сглаживающий подход, основанный на идеях метода максимального правдоподобия. В работе [18] рассматривается каскадный метод оценки изменяющихся по времени параметров. Этот метод оценивает параметры путем оптимизации определенного критерия, однако достижение глобального минимума проблематично с вычислительной точки зрения.

Другим подходом к оцениванию параметров обыкновенных дифференциальных уравнений (см. [19]) является двухэтапный метод, в котором на первом этапе, с использованием сглаживающих методов, по зашумленным данным оцениваются функция и ее производные, а затем на втором этапе строятся МНК-оценки параметров уравнения. Двухэтапный метод легко реализуется, однако он может быть статистически неэффективным, так как производные не могут быть точно оценены по зашумленным данным, особенно производные старших порядков.

Предложенная методика верификации основана на сопоставительном анализе параметров авторегрессионной модели и разностного уравнения. Полученные результаты показывают возможность использования такого подхода при моделировании зашумленных динамических объектов с распределенными переменными и обосновывают проведение дальнейших исследований в этом направлении.

Литература

[1] Захаров, Е.В., Дмитриева И.В., Орлик С.И.: Уравнения математической физики. Университетский учебник. М.: Издательский центр «Академия», 320 с. (2010)

[2] Dickey, D.A., Fuller, W.A.: Distribution of the Estimators for Autoregressive Time Series with a Unit Root. J. of the American Statistical Association, 74 (366), pp. 427-431 (1979). doi: 10.2307/2286348

[3] Носко, В.Н.: Эконометрика. Кн. 2. Ч. 3, 4. Учебник. М.: Изд. дом «Дело» РАНХиГС, 576 с. (2011)

[4] Мареев, В.В., Станкова, Е.Н.: Основы методов конечных разностей. СПб.: Изд-во С.-Петерб. ун-та, 64 с. (2012)

[5] Clements, M.P., Hendry, D.F.: Forecasting Non-stationary Economic Time series. Cambridge, Massachusetts, London: MIT Press, 262 p. (1999)

[6] Patterson, K.: An Introduction to Applied Econometrics: A Time Series Approach. New York: Palgrave, 832 p. (2000)

[7] Канторович, Г.Г.: Анализ временных рядов. Экономический журнал ВШЭ. (3), сс. 379-401 (2003)

[8] Хромов, С.П., Петросянц, М.А.: Метеорология и Климатология. Учебник. М.: Изд-во МГУ, 527 с. (2001)

[9] Матвеев, М.Г., Михайлов, В.В., Сирота Е.А.: Комбинированная прогностическая модель нестационарного многомерного временного ряда для построения пространственного профиля атмосферной температуры. Информационные технологии, 22 (2), сс. 89-94 (2016)

[10] NCEP/DOEAMIPIIReanalysis [Электронный ресурс]. http://www.cdc.noaa.gov/cdc/data.ncep.reanalysis2.html

[11] Бойков, И.В., Кривулин, Н.П.: Методы идентификации динамических систем. Программные системы: теория и приложения. 23 (5), С. 79-96 (2011)

[12] Bar, M., Hegger, R, Kantz, H.: Fitting Differential Equations to Space-time Dynamics. Physical Review, E. 59 (1), pp. 337-342 (1999). doi: 10.1103/PhysRevE.59.337

[13] Xun, X., Cao, J., Maity, J.: Parameter Estimation of Partial Differential Equation Models. J. of the American Statistical Association. 108 (503), pp. 1009-1020 (2013). doi: 10.1080/01621459.2013.794730

[14] Putter, H., Lange S.: A Bayesian Approach to Parameter Estimation in HIV Dynamical Models. Statistics in Medicine Heisterkamp, pp. 2199-2214 (2000)

[15] Huang, Y., Liu, D.: Hierarchical Bayesian Methods for Estimation of Parameters in a Longitudinal HIV Dynamic System, pp. 413-423 (2006)

[16] Ramsay, J. O., Hooker, G., Campbell, D.: Parameter Estimation for Differential Equations: a Generalized Smoothing Approach (with discussion). J. of the Royal Statistical Society, Series B, pp. 741-796 (2007)

[17] Cao, J., Huang, J., Wu, H.: Penalized Nonlinear Least Squares Estimation of Time-varying Parameters in Ordinary Differential Equations. J. of Computational and Graphical Statistics, pp. 42-56 (2012)

[18] Liang, H., Wu, H.: Parameter Estimation for Differential Equation Models Using a Framework of Measurement Error in Regression Models. J. of the American Statistical Association, pp. 1570-1583 (2008)

[19] Chen, J., Wu, H.: Efficient Local Estimation for Time-varying Coefficients in Deterministic Dynamic Models with Applications to HIV-1 Dynamics. J. of the American Statistical Association, pp. 369-384 (2008)

Представление новостных сюжетов с помощью событийных фотографий

© М.М. Постников © Б.В. Добров

Московский государственный университет имени М.В. Ломоносова
факультет вычислительной математики и кибернетики,
Москва, Россия

mihanlg@yandex.ru dobrov_bv@mail.ru

Аннотация. Рассмотрена задача аннотирования новостного сюжета изображениями, ассоциированными с конкретными текстами сюжета. Введено понятие «событийной фотографии», содержащей конкретную информацию, дополняющую текст сюжета. Для решения задачи применены нейронные сети с использованием переноса обучения (Inception v3) для специальной размеченной коллекции из 4114 изображений. Средняя точность полученных результатов составила более 94,7%.

Ключевые слова: событийная фотография, новостные иллюстрации, перенос обучения.

News Stories Representation Using Event Photos

© M.M. Postnikov © B.V. Dobrov

Lomonosov Moscow State University, Faculty of Computational Mathematics and Cybernetics,
Moscow, Russia

mihanlg@yandex.ru dobrov_bv@mail.ru

Abstract. The task of annotating a news story with images associated with specific texts is discussed in the article. The definition of "event photography" containing specific information supplementing text of a story is introduced. Neural networks (Inception v3) are used to solve a task for a special marked collection of 4114 images using the transfer learning method. The average precision of the results is more than 94.7%.

Keywords: event image, news illustration, transfer learning.

1 Введение

Распространение интернета, социальных сетей, развитие носимой электроники, внедрение хороших камер в каждый мобильный телефон – благодаря всем этим факторам, но не ограничиваясь ими, интернет становится главным источником новостей для современного человека, в то время как телевидение и печатные средства массовой информации постепенно уходят на второй план. С развитием технологий растут и скорость распространения информации, и ее количество.

Иллюстрации несут значительную часть информации, иногда даже большую, чем иллюстрируемый текст.

В данной работе рассмотрена задача определения изображений, «полезных» для понимания новостных сообщений, иллюстрируемых ими. Как известно, новостные сообщения прежде всего содержат информацию о некотором событии и должны отвечать на вопрос: «Что произошло?».

Рисунок 1 Пример несобытийной (слева) и событийной (справа) фотографии для новости с заголовком «в Краснодаре ГК СКИФ победил ставропольское «Динамо-Виктор» – 28:23»

Для ответа на общий вопрос обычно требуется ответить на совокупность частных вопросов: «Кто? Где? Когда? Каким образом?» и т. д. Важно научиться отличать полезные изображения от тех, которые не несут важной конкретной информации (см. Рис. 1).

В рамках данной работы *событийной фотографией* будем называть изображение, используемое для иллюстрации новости, для которого выполняются следующие требования:

а) изображение соответствует тексту новостной статьи;

б) изображение является фотографией с места событий или могло бы ей быть (подразумевается событие, описываемое в новостной статье).

Труды XIX Международной конференции «Аналитика и управление данными в областях с интенсивным использованием данных» (DAMDID/ RCDL'2017), Москва, Россия, 10–13 октября 2017 года

Например, фотографии с футбольного матча, места происшествия или встречи глав государств с соответствующими новостными текстами являются событийными. Если же на фотографии изображены логотип, рекламный баннер, вывеска, изображение взято из фотобанка или фотография не соответствует новости, тогда данная иллюстрация не будет считаться событийной.

Задача является актуальной при создании новостных агрегаторов по большому количеству источников.

Большой интерес для исследования представляют социальные сети – огромное количество свидетельств очевидцев в первую очередь публикуется в социальной сети, а затем уже может найти свое отражение в СМИ.

Идея работы состоит в том, чтобы попробовать выделить ключевые объекты на изображении, сопоставить их с текстом и достичь желаемого результата. Для распознавания объектов на изображении использованы сверточные нейронные сети.

В последнее время сверточные нейронные сети получили широкое распространение в обработке и классификации изображений, благодаря чему началась активная разработка фреймворков для удобной работы с ними (tensorflow [15], theano [17], keras [2] и др.), что снизило порог входа для применения данных технологий. Но обучение сложных моделей все еще отнимает значительное количество времени и средств. Например, обучать большую нейронную сеть для классификации на стандартном персональном компьютере без специальных компонентов можно и неделю, и месяц, а то и больше. И даже на мощной системе это занимает значительное количество времени [7, 9, 22].

Существуют подходы, например, метод *переноса обучения* [8], позволяющие существенно снизить временные издержки. Современные нейронные сети обработки изображений являются многослойными, последующие слои комбинируют признаки, выделенные на предыдущих уровнях. Начальные слои ответственны за выделение базовых примитивов изображения, следующие слои – за выделение типовых фигур как комбинаций базовых примитивов и т. д. Соответственно, можно попробовать взять сеть, обученную до некоторого уровня на одних коллекциях изображений, и дообучить ее на собственной коллекции (более подробно см. в разделе 5).

Также в работе исследована возможность улучшения качества выделения событийного изображения к новостному сообщению с использованием текста новости.

2 Постановка задачи

Формальная постановка задачи выглядит следующим образом.

Пусть T – множество новостных текстов, I – множество изображений, а $Y=\{0, 1\}$ – конечное множество оценок. Существует неизвестная целевая зависимость – отображение $F: T \times I \to Y$, значения которой известны только на объектах конечной обучающей выборки $\Omega = T^m \times I^m \times Y^m = \{(t_1, i_1, y_1), \dots, (t_m, i_m, y_m)\}$.

Будем также считать, что задана неотрицательная целочисленная функция потерь $L(\tilde{y}, y)$, которая показывает, насколько отличается предсказанное классификатором значение $\tilde{y}$ от истинного значения.

Обучающую выборку Ω разделим на две непересекающиеся коллекции:

- тренировочную (используется для обучения модели)
$$\Omega_{train} = (t_1, i_1, y_1), \dots, (t_n, i_n, y_n)\ ;$$
- тестовую (используется для оценивания модели)
$$\Omega_{test} = (t_{n+1}, i_{n+1}, y_{n+1}), \dots, (t_{|\Omega|}, i_{|\Omega|}, y_{|\Omega|}).$$

Задача классификации состоит в нахождении функции $F^* = F(t_j, i_j)$:

$$F^* = \operatorname{argmin}_F L(F(t_j, i_j), y_j), (t_j, i_j, y_j) \in \Omega_{test},$$

которая называется *классификатором*. Значение F^* может быть вещественным из диапазона [0;1], его можно считать вероятностью того, что изображение является событийным.

Решение задачи можно рассматривать как решение следующих трех подзадач.

2.1 Детектор объектов

На этапе обработки изображения построим модель, которая принимает на вход изображение, а возвращает вектор вероятностей присутствия определенных объектов на изображении.

Рисунок 2 Пример изображения для детектирования объектов

Например, пусть на вход подается следующее изображение (см. Рис. 2). Пусть рассматривается присутствие следующих классов: *мотоцикл, автомобиль, человек, домашнее растение, велосипед, автобус, поезд, птица, лодка, лошадь, самолет, бутылка, телевизор, кресло, собака, кот, стол, кровать, корова, овца*. Результатом работы детектора может быть следующий вектор вероятностей: [0.9984, 0.4156, 0.0144, 0.006, 0.003, 0.0009, 0.0008, 0.0007, 0.0005, 0.0004, 0.0001, 0.0001, 0, …, 0]. Значение на позиции i представляет собой вероятность присутствия объекта класса i из списка. В данном случае *мотоцикл* присутствует на изображении с вероятностью 99,9%, а *автомобиль* – с вероятностью 41,6%.

2.2 Векторизация текста

Для обработки текста будем обучать модель, которая создает векторное представление новостного текста, поданного на вход классификатору.

2.3 Модель согласованности

Финальный этап в работе классификатора объединяет в себе итоги предыдущих подзадач. Модель, используемая на этом этапе, принимает на вход два вектора – вектор, полученный в результате обработки изображения, и вектор, полученный в результате обработки текста. Промежуточные представления векторов объединяются в единый вектор, а на выходе модели получается число из интервала [0;1] – вероятность того, что входное изображение является событийным для входной новостной статьи.

3 Обзор

Между изображениями и текстовой информацией, которая может быть ассоциирована изображению, существуют достаточно сложные взаимосвязи в зависимости от контекста и решаемых задач.

В настоящей статье рассматривается задача выбора лучшего изображения среди возможных для новостного текста с использованием информации об объектах, которые можно выделить на изображении, а также информации о связи текста с выделенными объектами.

Известны похожие постановки задач для решения проблем описания изображений текстом (Image Caption), поиска изображений по текстовому запросу (Visual Question Answering). Одним из направлений решения задач, возникающих в данных областях, является определение типа события, отображаемого на картинке/фотографии [1]. Для этой цели создаются коллекции изображений [21], аннотируемых либо свободным описанием несколькими экспертами, либо тегами [3].

К сожалению, существует несколько фундаментальных проблем описания изображений текстом. Имеет место существенный семантический разрыв между семантикой изображения и семантикой текста – обычно слишком много деталей опущено. Эксперты, описывающие изображения, могут по-разному их понимать, в том числе в силу разного жизненного опыта. Кто-то видит просто мужчину, а кто-то – известного актера, кто-то видит просто темный диск, а кто-то – грампластинку, и т. д. Кто-то может не описать тот или иной фрагмент изображения, так как он показался ему неинтересным. Существующие иерархии концептов для описания изображений пока существенно неполны.

В результате в работе [18] отмечено, что только 20% проанализированных авторами описаний изображений не содержат ошибок, при этом 26% описаний по мнению авторов не релевантны изображениям.

Отметим также, что часто банки изображений формируются из всех фотографий, сделанных во время события. Однако некоторые фотографии могут содержать не главные объекты, но окружение, которое, вообще говоря, не является специфичным именно для конкретного события (типа события).

В статье [1] приведены данные, что качество распознавания конкретных событий по размеченным коллекциям в настоящее время имеет следующие характерные оценки (на коллекции WIDER [21], 60 классов, 60 000 изображений): 42% корректных ответов среди первых, 60% – среди первых пяти.

Таким образом, пока нет возможности с высокой степенью уверенности опереться на методы определения типа события по изображению, использовать методы порождения описания изображений. При этом в [19] определена характеристика «важности» того или иного типа объекта для описания того или иного типа события, например, изображение местных достопримечательностей для альбома о путешествии или изображения невесты и жениха для фотоальбома о свадьбе. Авторы [19] предлагают выделять наиболее важные объекты путем выявления среднего по большому количеству изображений о событиях одного типа.

4 Модели

Для построения детектора объектов на изображении используется комбинация из двух моделей. Первая из них – обученная на большом объеме изображений сверточная нейросеть, используемая для извлечения вектор-признаков с изображения. Вторая модель – это основной классификатор, который преобразует полученные вектор-признаки первой модели в нужные нам «вероятностные» признаки.

Нейронная сеть – широко используемый метод машинного обучения, показывающий отличные результаты в анализе изображений, текстов, распознавании речи и других областях. В последнее время сверточные нейронные сети получили большое распространение, и эта область машинного обучения сейчас активно развивается.

На вход первой нейронной сети подается изображение, на выходе получается некоторый вектор-признак, который далее подается на вход основному классификатору.

В качестве основного классификатора рассмотрим следующие модели:

- логистическая регрессия;
- градиентный бустинг;
- нейронная сеть.

4.1 Логистическая регрессия

Логистическая регрессия – это линейный алгоритм классификации с логистической функцией потерь. Часто эта модель используется в качестве отправной точки (baseline).

$$a(x, w) = \text{sign}(\sum_{i=1}^{n} w_j f_j(x) - w_0) = \text{sign}\langle w, x\rangle,$$

где w_j – вес j-го признака, w_0 – порог принятия

решения, $w = (w_0, w_1, \dots, w_n)$ – вектор весов, $\langle w, x \rangle$ – скалярное произведение признакового описания объекта на вектор весов. Считается, что $f_0(x) = -1$,

$$L(w) = \sum_{i=1}^{m} \ln(1 + e\exp^{-y_i * \langle x_i, w \rangle}) \to \min_w.$$

Логистическая регрессия – статистическая модель, которая используется для предсказания вероятности возникновения некоторого события по значениям множества признаков:

$P\{y|x\} = \sigma(y\langle x, w\rangle), \quad \sigma(z) = 1/(1 + \exp^{-z})$.

Модели обучаются на вектор-признаках – выходе первой нейронной сети.

Как основной классификатор рассматривается набор моделей логистических регрессий – для каждого выделенного класса используется своя модель. Реализация логистической регрессии берется из библиотеки sklearn с параметрами по умолчанию [11, 12]. Результаты этих моделей затем объединяются в один вектор.

4.2 Градиентный бустинг

Градиентный бустинг – метод машинного обучения, основанный на ансамбле деревьев решений, считающийся одним из наиболее эффективных методов (с точки зрения качества классификации) и обладающий хорошей обобщающей способностью.

Градиентный бустинг, как и любой бустинг-алгоритм, последовательно строит базовые модели так, что каждая следующая улучшает качество всего ансамбля. Градиентный бустинг деревьев решений строит модель в виде суммы деревьев:

$$f(x) = h_0 + \sum_{j=1}^{M} b_j h(x; a_m),$$

где h_0 – некоторое начальное приближение, $b_j \in R$ – параметр, регулирующий скорость обучения и влияние отдельных деревьев на всю модель, $h_j(x; a_n)$ – базовый алгоритм с вектором параметров a_n.

$L = \sum_{i=1}^{N} L(y_i, f_j(x_i)) \to \min_{a_i, b_i}$ – некоторая функция потерь.

Модели обучаются на вектор-признаках – выходе первой нейронной сети.

Аналогично классификатору, использующему логистическую регрессию, рассматривается набор моделей градиентного бустинга – по одной на каждый класс. Реализация градиентного бустинга берется из библиотеки sklearn с параметрами по умолчанию [10, 12]. Результаты этих моделей так же объединяются в один вектор.

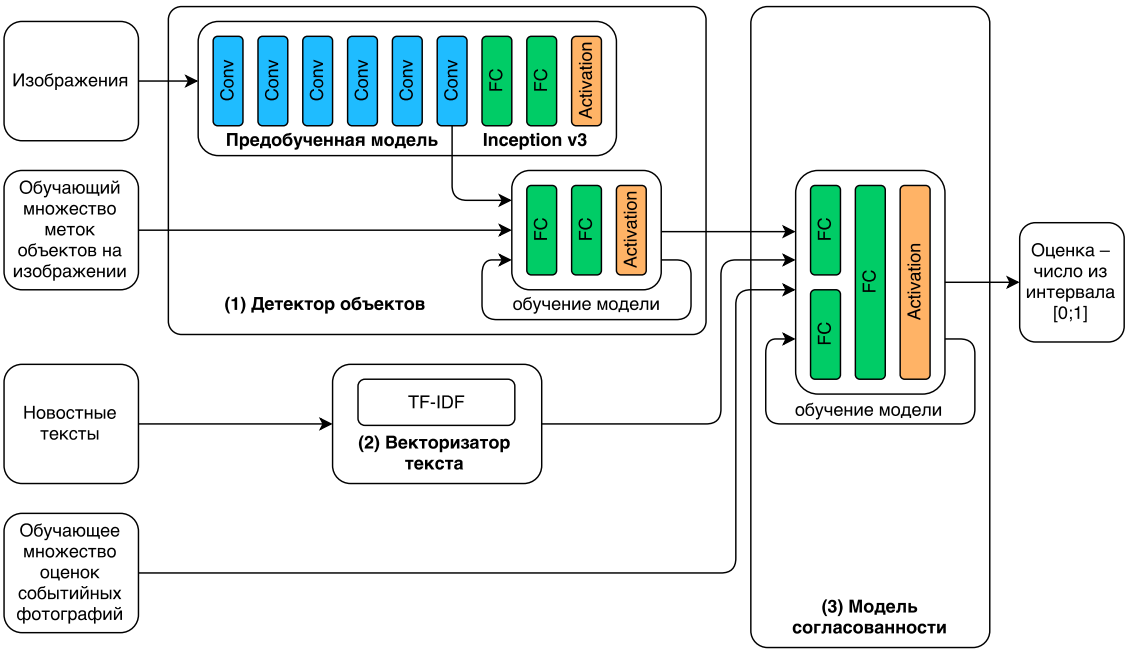


Рисунок 3 Схема потока данных обучения моделей

4.3 Нейронная сеть

Рассмотрим нейросеть, на вход которой подается вектор-признак с первой нейросети, а на выходе получается вектор, описывающий вероятность присутствия классов на изображении.

В проведенном исследовании использована следующая архитектура нейронной сети (см. Рис. 3, детектор объектов). Такая архитектура используется в финальных слоях модели VGG-16 [9], которые идут сразу же за сверточными. В оригинале последний слой – Softmax, который не очень подходит для нашей задачи (при повышении оценки для одного класса все остальные занижаются). Нами последний слой был заменен на Sigmoid, так как решается задача многозначной классификации.

Для обучения нейронной сети использована следующая функция потерь:

$$L(\tilde{y}, y) = -(y \log y + (1 - y)\log(1 - \tilde{y}))$$

– бинарная кросс-энтропия.

5 Методы

5.1 Обработка изображения

Для представления изображения в виде вектора используются модели, описанные в п. 4.1. Для данного преобразования применяется метод, называемый переносом обучения (transfer learning).

Перенос обучения – метод, позволяющий применить знания, полученные в процессе решения одной задачи, для решения другой схожей задачи. Например, можно взять уже предобученную на большом объеме данных нейронную сеть и дообучить ее на своих данных. Применение данного метода обычно позволяет сэкономить большое количество ресурсов (как времени, так и вычислительных ресурсов). В данной работе в качестве предобученной модели использована Inception v3, обученная для ImageNet Large Visual Recognition Challenge на данных 2012 года [14].

В случае обработки изображений берется первая модель (предобученная нейросеть), на вход которой подается изображение. Затем из этой сети извлекается некоторый слой, который и будет являться промежуточным векторным представлением нашего изображения. Данный слой обычно содержит большое количество признаков, помогающих решать задачу классификации. Далее этот слой подается на вход уже второй модели (линейной регрессии, градиентному бустингу или другой нейронной сети). На выходе мы получаем вектор вероятностей присутствия объектов для каждого из классов.

5.2 Обработка текста

Для представления текста в векторном виде используется TF-IDF. Для каждого документа из коллекции его исходный текст токенизируется, а токены приводятся в начальную форму. Затем считается подокументная частотность преобразованных токенов и вычисляется TF-IDF вектор для каждого документа.

5.3 Объединение текста и изображений

Результаты обработки коллекции новостей с изображениями по пп. 5.1 и 5.2 подаются на вход нейросети, которая комбинирует полученную информацию с двух входов и возвращает число в диапазоне от 0 до 1 – вероятность того, что изображение подходит для иллюстрации текста. Нами проверено, влияет ли использование текстовых признаков на качество определения фотографии как событийной или же достаточно использования только признаков, выделенных на изображении.

6 Исходные данные

В качестве данных для обучения и отладки моделей были использованы как готовые наборы данных, так и специально собранные для решения поставленной задачи.

6.1 CIFAR-10

CIFAR-10 – это коллекция размеченных изображений, взятых из другого набора данных под названием «80 million tiny images» [16].

Описание коллекции:

- 60000 размеченных изображений;
- 10 классов, 6000 изображений на класс (*самолет, автомобиль, птица, кошка, олень, собака, лягушка, лошадь, корабль, грузовик*);
- размер изображений фиксированный, 32x32;
- один класс на одном изображении.

6.2 Pascal VOC2012

Pascal VOC2012 – это коллекция размеченных изображений, которые были собраны для соревнования по распознаванию и классификации объектов [5].

Описание коллекции:

- 11540 размеченных цветных изображений;
- 20 классов, в среднем по 577 изображений на класс (*мотоцикл, автомобиль, человек, домашнее растение, велосипед, автобус, поезд, птица, лодка, лошадь, самолет, бутылка, телевизор, кресло, собака, кот, стол, кровать, корова, овца*);
- размер изображений не фиксирован, но максимальная длина сторон – 500 пикселей;
- не менее одного класса на изображении.

6.3 Коллекция изображений на базе ImageNet

Для обучения детектора объектов нужно просмотреть новостные иллюстрации, понять, какие объекты там чаще всего встречаются, и собрать собственную коллекцию для обучения. Нами выделено 38 классов объектов, которые чаще всего встречаются в новостных иллюстрациях, и для них была собрана коллекция изображений на базе ImageNet [6].

Описание коллекции:

- 62357 размеченных цветных изображений;
- 38 классов, в среднем по 1640 изображений на класс (*воздушная техника, животное, баннер, лодка, здание, церковь, концерт, конструкция, толпа, документ, электронное устройство, огонь/дым, флаг, еда, в помещении, военная воздушная техника, военный транспорт, гора, нефтегазовые строения, картина, человек, растение, общественный транспорт, дорога, корабль, снаружи, солдат, космический корабль, спикер, транспорт специальных служб, спорт, деловой костюм, телекамера, служебная форма, транспорт, военный корабль, вода, оружие);*
- размер изображений не фиксирован;
- не менее одного класса на изображении.

6.4 Коллекция новостей

В качестве обучающей коллекции для определения событийной фотографии был собран набор новостей, содержащий 4114 примеров. В результате разметки получилось 3100 позитивных и 1014 негативных примеров событийных фотографий.

Таблица 1 Значение AP для 4 моделей на наборе данных Pascal VOC2012

	Самолет	Велосипед	Птица	Лодка	Бутылка	Автобус	Автомобиль	Кот	Кресло	Корова	mAP
GBC	0,9664	0,7670	0,8934	0,8090	0,5034	0,8747	0,8079	0,9050	0,7331	0,7241	
LR	**0,9938**	**0,8769**	0,9117	0,8795	0,5294	**0,9105**	0,8197	0,9140	0,7115	0,7402	
NN	0,9628	0,8738	0,9140	0,8885	**0,6142**	0,9089	**0,8229**	**0,9148**	**0,7796**	**0,7466**	
*HCP**	*0,9750*	*0,8430*	*0,9300*	*0,8940*	*0,6250*	*0,9020*	*0,8460*	*0,9480*	*0,6970*	*0,9020*	
	Стол	Собака	Лошадь	Мотоцикл	Человек	Растение	Овца	Кровать	Поезд	Телевизор	
GBC	0,7467	0,8729	0,8832	0,8748	0,8905	0,5145	0,7992	0,6076	0,8888	0,7276	0,7895
LR	0,6350	0,8995	**0,9458**	0,9010	0,9010	0,5017	0,8122	0,6900	0,9034	0,7464	0,8112
NN	**0,7480**	**0,9188**	0,9099	**0,9212**	**0,9062**	**0,5559**	**0,8272**	**0,7716**	**0,9050**	**0,7847**	**0,8337**
*HCP**	*0,7410*	*0,9340*	*0,9370*	*0,8880*	*0,9330*	*0,5970*	*0,9030*	*0,6180*	*0,9440*	*0,7800*	*0,8420*

7 Эксперименты

7.1 Выбор модели для обработки изображений

В качестве основной метрики была использована AP (average precision), определенная в статье [4] и вычисляемая по следующей формуле:

$$AP = \frac{1}{11} \sum_{r \in \{0, 0.1, \ldots, 1\}} p_{interp}(r),$$

$p_{interp}(r) = \max_{\tilde{r}:\tilde{r} \geq r} p(\tilde{r})$ – интерполяция точности, где $p(\tilde{r})$ – это измеренное значение точности для значения полноты $\tilde{r}$, p(x) – кривая «точность–полнота».

Сравнения качества моделей на наборе данных Pascal VOC2012 отображено в таблице1. Здесь также приведены значения AP для модели HCP-2000C [20] на конкурсном тестовом множестве (не на том, который был использован для сравнения моделей 1–3).

Из полученных оценок можно сделать вывод, что нейронная сеть с текущими параметрами лучше подходит для решения поставленной задачи, в дальнейшем будем рассматривать ее как основную модель.

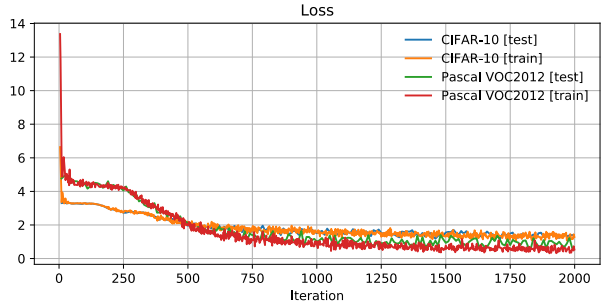


Рисунок 4 Зависимость значения функции потерь от итерации

На графике (см. Рис. 4) можно наблюдать, как сходится функция потерь нейросетевой модели на различных наборах данных. Поведение функций довольно похожее, но на Pascal VOC2012 при более медленной сходимости достигается лучшее качество. Графики MAE (средней абсолютной ошибки) ведут себя одинаковым образом (см. Рис. 5).

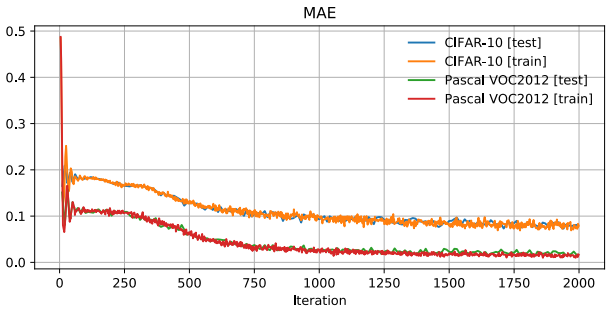


Рисунок 5 Зависимость значения MAE от итерации

7.2 Обучение модели на собственном наборе данных

Убедившись, что модель работает и показывает хорошие результаты на готовых наборах данных, нужно перейти к следующему этапу – обучению модели на собственной коллекции.

7.3 Применение моделей к новостям

Для этого обучается модель согласованности, которая по входным данным определяет, является изображение событийным или нет.

Обучим две модели, одна из которых принимает на вход одно лишь векторное представление изображения, а другая принимает на вход, помимо прочего, векторное представление соответствующего новостного текста. Во второй сети каждый из двух векторов входа преобразуется в вектор общей длины, затем конструируется новый вектор, получающийся конкатенацией поэлементного умножения и поэлементного сложения предыдущих слоев. Финальный слой

каждой сети – Softmax. На выходе нейросети получаем два значения p_1, p_2 в интервале [0;1], первое из которых – вероятность того, что изображение не является событийным для этой фотографии, а второе – что является ($p_1 + p_2 = 1$). Для обучения нейронной сети использована следующая функция потерь: $L(\tilde{y}, y) = -(y \log \tilde{y} + (1-y)\log(1-\tilde{y}))$ – софтмакс кросс-энтропия.

На Рис. 6 показаны графики значения функции потерь для каждой из двух моделей. Отметим, что модель, использующая текст, имеет склонность к переобучению после ~1500 итераций.

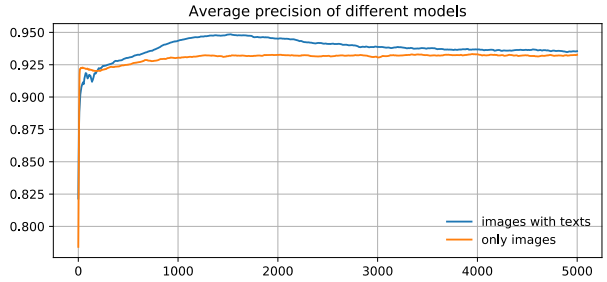


Рисунок 6 Зависимость значений средней точности для различных моделей на тестовых данных

7.4 Результаты

Следующие шаги после обучения детектора – его применение к новостным изображениям, получение векторного представления изображений и перевод текстов в векторную форму. Примеры работы программы изображены на Рис. 7 и 8.

Рисунок 7 Пример работы программы

Рисунок 8 Пример работы программы

8 Интерпретация результатов

В работе исследован метод ранжирования изображений для иллюстрации новостного сюжета, а именно, выявления изображений, которые с большей вероятностью содержат информацию, дополняющую текстовое сообщение. Представлен метод с использованием переноса обучения результатов Inception v3, когда несколько последних слоев обученной нейронной сети заменяются специфическим классификатором для исследуемой коллекции изображений.

В проведенных экспериментах специфический классификатор на основе нейронных сетей несколько превзошел логистическую регрессию и градиентный бустинг (однако для практических целей данные методы также можно использовать).

На коллекции из 4114 изображений (из них 3100 событийных), размеченной одним из авторов, достигнут результат 93,2% средней точности при обучении только по изображениям и 94,7% при использовании текстовой информации.

Целью дальнейших исследований являются применение более сложных и современных моделей классификации, введение дополнительных признаков, выделенных на изображениях, оценка применимости данной работы на таких источниках новостей, как социальные сети, улучшение и расширение собранных коллекций.

Литература

[1] Ahsan, U., Sun, C., Hays, J., Essa, I.: Complex Event Recognition from Images with Few Training Examples. Applications of Computer Vision (WACV), 2017 IEEE Winter Conference on, pp. 669-678 (2017)

[2] Chollet, F. and others: Keras. https://github.com/fchollet/keras

[3] Cui, Y., Liu, D., Chen, J., Chang, S.F.: Building a Large Concept Bank for Representing Events in Video. arXiv preprint arXiv:1403.7591 (2014)

[4] Everingham, M., Van Gool, L., Williams, C.K.I., Winn, J., Zisserman, A.: The PASCAL Visual Object Classes (VOC) Challenge: A Retrospective. Int. J. of Computer Vision, 111 (1), pp. 98-136 (2015)

[5] Everingham, M., Winn, J.: The PASCAL Visual Object Classes Challenge 2012 (VOC2012) Development Kit (2012)

[6] ImageNet. http://image-net.org/index

[7] Krizhevsky, A., Sutskever, I., Hinton, G.E.: ImageNet Classification with Deep Convolutional Neural Networks. Advances in Neural Information Processing Systems, pp. 1097-1105 (2012)

[8] Oquab, M., Bottou, L., Laptev, I., Sivic, J.: Learning and Transferring Mid-Level Image Representations using Convolutional Neural. Networks. M.: CVF (2014)

[9] Simonyan, K., Zisserman, A.: Very Deep 5Convolutional Networks for Large-Scale Image Recognition. arXiv preprint arXiv:1409.1556 (2014)

[10] Sklearn, GradientBoostingClassifier. http://scikit-learn.org/stable/modules/generated/sklearn.ensemble.GradientBoostingClassifier.html

[11] Sklearn, LogisticRegression. http://scikit-learn.org/stable/modules/generated/sklearn.linear_model.LogisticRegression.html

[12] Sklearn. OneVsRestClassifier. http://scikit-learn.org/stable/modules/generated/sklearn.multiclass.OneVsRestClassifier.html

[13] Srivastava, N., Hinton, G., Krizhevsky, A., Sutskever, I., Salakhutdinov, R.: Dropout: A Simple Way to Prevent Neural Networks from Overfitting. J. of Machine Learning Research, 15 (1), pp. 1929-1958 (2014)

[14] Szegedy, C., Vanhoucke, V., Ioffe, S., Wojna, Z., Shlens, J.: Rethinking the Inception Architecture for Computer Vision. Proc. of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 2818-2826 (2016)

[15] TensorFlow: Large-scale machine learning on heterogeneous systems. http://tensorflow.org (2015)

[16] The CIFAR-10 dataset. https://www.cs.toronto.edu/~kriz/cifar.html

[17] Theano Development Team. Theano: A Python Framework for Fast Computation of Mathematical Expressions. arXiv preprint arXiv:1605.02688 (2016)

[18] van Mitenburg, E., Elliot, D.: Room for Improvement in Automatic Image Description: an Error Analysis. arXiv preprint arXiv:1704.04198 (2017)

[19] Wang, Y., Lin, Z., Shen, X., Mech, R., Miller, G., Cottrell, G.W.: Event-specific Image Importance. Proc. of the IEEE Conf. on Computer Vision and Pattern Recognition, pp. 4810-4819 (2016)

[20] Wei, Y., Xia, W., Huang, J., Ni, B., Dong, J., Zhao, Y., Yan, S.: CNN: Single-label to Multi-label. arXiv preprint arXiv:1406.5726 (2014)

[21] Yang, S., Luo, P., Loy, C.C., Tang, X.: Wider Face: A Face Detection Benchmark. Proc. of the IEEE Conf. on Computer Vision and Pattern Recognition, pp. 5525-5533 (2016)

[22] Zeiler, M.D., Fergus, R.: Visualizing and Understanding Convolutional Networks. European Conf. on Computer Vision. Springer, Cham, pp. 818-833 (2014)

Метод прогнозирования развития ситуаций на основе обнаружения событий в потоке текстовых документов

© А.М. Андреев © Д.В. Березкин © И.А. Козлов

Московский государственный технический университет им. Н. Э. Баумана, Москва

arkandreev@gmail.com berezkind@bmstu.ru kozlovilya89@gmail.com

Аннотация. Рассмотрен метод автоматизированного прогнозирования развития ситуаций на основе обнаружения событий в потоке текстовых документов. Описаны существующие подходы к анализу ситуаций, выявлены их преимущества и недостатки с точки зрения специфики решаемой задачи. Предложен метод формирования сценариев развития ситуаций на основе принципа исторической аналогии, учитывающий динамику развития ситуаций. Этот метод позволяет оценивать вероятность реализации сформированных сценариев с помощью логистической регрессии. Представлен метод выделения оптимистического и пессимистического сценариев на основе метода анализа иерархий. Описан способ снабжения сценариев предложениями для лиц, принимающих решения. Представлены результаты экспериментальной оценки качества разработанного метода.

Ключевые слова: ситуационный анализ, прогнозирование, сценарный анализ, система поддержки принятия решений, аналогия, анализ текстового потока.

Method for Forecasting of Situations Development Based on Event Detection in Text Stream

© Ark Andreev © Dmitry Berezkin © Ilya Kozlov

Bauman Moscow State Technical University, Moscow

arkandreev@gmail.com berezkind@bmstu.ru kozlovilya89@gmail.com

Abstract. The article deals with the problem of automated forecasting of situations development based on event detection in a stream of text documents. Existing methods of situational analysis are analyzed and their advantages and disadvantages in view of the specifics of the task are determined. A method for generation of possible scenarios of situations development is described. The method generates scenarios on the principle of historical analogy, taking into account the dynamics of situation development. The probability of the generated scenarios' implementation is estimated via logistic regression. A method for the optimistic and the pessimistic scenario identification based on analytic hierarchy process is proposed. A way to supplement scenarios with recommendations for decision-makers is described. The results of experimental evaluation of the developed method's quality are presented.

Keywords: situational analysis, forecasting, scenario analysis, decision support system, analogy, text stream analysis.

1 Введение

В настоящий момент большое количество данных, обрабатываемых современными информационными системами (ИС), имеет форму информационных потоков: новые информационные сообщения постоянно поступают из источников и должны обрабатываться ИС с минимальной задержкой. Как правило, информация в потоке представлена в неструктурированном виде, в частности, в форме текста. Так, форму текстовых потоков имеют сообщения пользователей в социальных сетях, новости СМИ, официальные заявления органов власти.

Динамический характер текстовых потоков делает их важным средством информационной поддержки для людей, которым требуется принимать управленческие решения в режиме реального времени в условиях меняющейся обстановки. Задачи своевременного обнаружения проблемной ситуации, отслеживания её развития и оперативного принятия решений по управлению развитием ситуации возникают в различных сферах – политической, социальной, военной, экономической.

Труды XIX Международной конференции «Аналитика и управление данными в областях с интенсивным использованием данных» (DAMDID/ RCDL'2017), Москва, Россия, 10–13 октября 2017 года

Анализ текстового потока позволяет осуществлять мониторинг интересующих пользователей тем, т. е. обнаруживать возникновение важных событий, относящихся к тем или иным явлениям или объектам [5]. Обнаруживаемые события отражают развитие различных ситуаций с течением времени. Однако для принятия наилучших решений необходимо также определять возможные варианты дальнейшего развития этих ситуаций – это позволяет на основе полученного прогноза предпринимать определенные шаги, направленные на изменение ситуации в нужную сторону.

В статье предложено решение задачи автоматизированного прогнозирования развития ситуаций на основе анализа потока текстовых сообщений.

2 Постановка задачи

2.1 Функционирование системы мониторинга развития ситуаций

В [5] предложено решение задачи мониторинга тем на основе обнаружения событий, релевантных заданным темам, в потоке текстовых документов. Под событием понимается некоторое изменение, произошедшее в реальном мире и отраженное в текстовом потоке. Обнаружение событий рассматривается как задача кластеризации, заключающаяся в разбиении текстового потока на группы документов, описывающих различные события. Для этого каждый документ представляется многокомпонентной моделью, компоненты которой описывают содержание, структуру и метаданные документа: $d_i = (d_i^w, d_i^{tw}, d_i^c, d_i^p, d_i^n, d_i^{dt}, d_i^e, d_i^g, d_i^t)$. В частности, текстовое содержание документа представлено вектором $d_i^w = (w_i^1, w_i^2, \dots, w_i^{N^w})$, каждый элемент которого w_i^k отражает значимость k-го терма в контексте документа и рассчитывается с помощью метода TF-IDF. Каждое событие также описано многокритериальной моделью, компоненты которой формируются на основе документов, относящихся к событию: $\varepsilon_j = (\varepsilon_j^w, \varepsilon_j^{tw}, \varepsilon_j^c, \varepsilon_j^p, \varepsilon_j^n, \varepsilon_j^{dt}, \varepsilon_j^e, \varepsilon_j^g, \varepsilon_j^t)$. Объединение документов в группы выполняется с помощью алгоритма инкрементальной кластеризации, в основе которого лежит покомпонентное сопоставление каждого документа с ранее обнаруженными событиями. Более подробно модели документа и события, а также метод обнаружения событий описаны в [5].

Метод позволяет работать с документами на произвольном языке при наличии подготовленных экспертами тематических запросов, а также словарей имен персон, названий организаций и географических наименований на соответствующем языке. Для повышения качества обнаружения событий могут быть использованы наработки авторов в области морфологического [6], синтаксического [4] и семантического [7] анализа текстов. Для отслеживания изменения обстановки с течением времени необходимо формировать ситуации – цепочки взаимосвязанных событий, отражающие развитие тех или иных процессов. Для этого из множества обнаруженных событий выделяют пары взаимосвязанных событий $p_{ij} = (\varepsilon_i, \varepsilon_j)$, потенциально принадлежащих одной ситуации. На основе формирования таких пар выполняется построение ситуационного графа $G = (E, P)$. В этом графе узлы $E = \{\varepsilon_i\}$ соответствуют событиям, а ребра $P = \{p_{ij}\}$ – выделенным парам (каждое ребро является ориентированным и направлено к более позднему событию пары). Любой путь в этом графе является потенциальной ситуацией $s = (\varepsilon_s^1, \varepsilon_s^2, \dots, \varepsilon_s^n)$.

На Рис. 1 представлен пример ситуации, представляющей собой последовательность из четырех взаимосвязанных событий.

2.2 Особенности решаемой задачи

Прогнозирование заключается в построении возможных сценариев развития ситуации. Каждый сценарий представляет собой потенциальное продолжение текущей ситуации, т. е. цепочку событий, которые могут наступить в будущем. Для эффективного использования результатов прогнозирования из множества сформированных сценариев необходимо выделить три варианта, представляющих наибольший интерес для лиц, принимающих решения (ЛПР), – пессимистический, оптимистический и наиболее вероятный. На основе результатов прогнозирования должны приниматься решения по управлению ситуацией. Поэтому помимо сформированных сценариев пользователю должны предлагаться предложения по действиям, которые необходимо предпринять для способствования развитию ситуации по наиболее благоприятному сценарию.

3 Обзор существующих подходов к анализу ситуаций

В некоторых работах, посвященных ситуационному анализу, ситуации описываются совокупностями определенных числовых показателей [13]. Для прогнозирования в этом случае могут использоваться методы анализа временных рядов и методы регрессионного анализа. Такие подходы не могут быть использованы для анализа развития ситуаций на основе текстового потока, поскольку требуемый результат прогнозирования имеет качественный, а не количественный характер.

В ряде работ предложены подходы к формированию сценариев на основе когнитивных карт и знаковых орграфов [14, 16]. В них ситуация представляется как граф, узлы которого соответствуют факторам ситуации, а ребра отражают влияние факторов друг на друга. Прогнозирование заключается в оценке будущих значений факторов путем моделирования изменения ситуации с учетом различных управляющих воздействий. Построение описания ситуации в виде когнитивной карты выполняется экспертом, поэтому такие подходы неприменимы для автоматического формирования сценариев развития ситуаций.

Ситуация: Тестирование беспилотных такси Uber

Компания Uber запустила беспилотное такси в США

Имя документа	Дата публикации	Время публикации	Источник
Компания Uber запустила беспилотное такси в США	14.09.2016	14:41:39	ТАСС
Uber запустил первые беспилотные такси в США	14.09.2016	17:19:37	РБК
Беспилотные такси выехали на дороги	14.09.2016	19:40:00	Комсомольская Правда

Власти Калифорнии требуют от Uber прекратить использование беспилотных такси

Имя документа	Дата публикации	Время публикации	Источник
Власти Калифорнии требуют, чтобы Uber свернул сервис беспилотного такси в Сан-Франциско	15.12.2016	07:27:16	ТАСС
Власти Калифорнии требуют от Uber прекратить использование беспилотных такси	15.12.2016	07:55:00	Коммерсант
Власти Калифорнии потребовали прекратить эксперимент Uber с беспилотными такси	15.12.2016	09:15:00	Интерфакс
В Калифорнии потребовали ликвидировать сервис беспилотного такси Uber	15.12.2016	09:28:00	Комсомольская Правда

Власти вынудили Uber свернуть онлайн-вызов такси с автопилотом в Сан-Франциско

Имя документа	Дата публикации	Время публикации	Источник
Власти вынудили Uber свернуть онлайн-вызов такси с автопилотом в Сан-Франциско	22.12.2016	06:01:00	ТАСС
Uber приостановил испытания беспилотных такси в Калифорнии	22.12.2016	10:59:00	Интерфакс

Uber перенесла беспилотные такси в Аризону

Имя документа	Дата публикации	Время публикации	Источник
Uber перенес испытания беспилотных такси в Аризону	23.12.2016	19:51:00	Интерфакс
Uber перенесла беспилотные такси в Аризону	23.12.2016	20:10:00	Вести Экономика
Uber после неудачи в Калифорнии протестирует сервис беспилотного такси в Аризоне	24.12.2016	07:01:18	ТАСС

Рисунок 1 Пример выявления событий и формирования ситуации

Многие работы используют принцип аналогии – прогнозирование дальнейшего развития ситуации и формирование предложений по управляющим действиям основано на поиске аналогичных ситуаций, имевших место в прошлом. В работах, базирующихся на принципе аналогии, используются различные подходы к представлению ситуаций.

В [10] ситуация представляется фрагментом семантической сети, содержащим объекты и их отношения в рамках ситуации. Получить такое представление автоматически возможно лишь для определенных предметных областей, поэтому такой подход нельзя использовать для прогнозирования развития произвольных ситуаций.

В ряде работ предложено описание ситуаций в виде набора или вектора параметров с определенными значениями [1, 12]. Для сравнения ситуаций с целью определения аналогии используются евклидово расстояние, манхэттенская метрика, расстояние Чебышева, мера Хэмминга, косинусная мера и другие меры близости. Недостаток данных подходов заключается в статическом описании ситуаций – при определении близости между ситуациями не учитывается сходство динамики их развития.

Для учета динамики можно использовать описание эталонной ситуации в виде графа или автомата [3, 9, 11, 15, 18], пути в котором отражают различные варианты развития ситуации. Все эти подходы позволяют применять лишь принцип строгой аналогии: анализируемая текущая ситуация должна точно соответствовать некоторому пути в графе, построенном экспертом. Однако цепочка событий, автоматически построенная при анализе текстового потока, не всегда точно соответствует эталону – в ней могут содержаться дополнительные события или, напротив, отсутствовать какие-либо события из графа.

Подход на основе нестрогой аналогии предложен в [8]. Ситуации представляются цепочками событий, близость между ними определяется с помощью модифицированного расстояния Левенштейна. Но этот подход требует выделения для каждого события объекта и субъекта, что не может быть сделано автоматически для произвольных текстовых сообщений. Кроме того, результат определения аналогов текущей ситуации в названной работе используется лишь для отнесения этой ситуации к одному из заданных классов.

4 Предлагаемый метод прогнозирования развития ситуаций

Обнаружение для текущей последовательности s_c цепочки-аналога s_e позволяет не только определить вероятный итог развития ситуации (как предлагается в [8]), но и объяснить, какие события могут привести к этому итогу. Такой прогноз можно получить, если обнаружено сходство всей текущей последовательности с начальной частью $st(s_e, s_c)$ цепочки-аналога. В этом случае можно предположить, что в будущем наступят события, аналогичные тем, которые составляют заключительную часть цепочки-аналога $fin(s_e, s_c)$. Таким образом, эту заключительную часть можно рассматривать как возможный сценарий дальнейшего развития текущей ситуации.

Для выполнения сопоставления необходимо наличие базы ситуаций-эталонов $S_e = \{s_e^i\}$. Такие эталоны отбираются экспертами в зависимости от

задачи, для которой используется система мониторинга. Так, для анализа ситуации, связанной с тестированием беспилотных такси (рис. 1), использовалась база эталонных ситуаций, отражающих развитие различных технологий в прошлом.

Поскольку текущие ситуации представляют собой пути в ситуационном графе, процесс прогнозирования состоит из следующих этапов:

1. При появлении в ситуационном графе нового события ε_c (либо при изменении существующего события) осуществляется поиск аналогичных ему событий, принадлежащих эталонным ситуациям.
2. При нахождении эталонного события $\varepsilon_e \in s_e$, аналогичного событию ε_c, выполняется попытка выделить в графе цепочку событий s_c (текущую ситуацию), которая содержит событие ε_c и имеет максимальное сходство с начальной частью $st(s_e, s_c)$ последовательности s_e. Если s_c является аналогом s_e, то заключительная часть эталонной ситуации $fin(s_e, s_c)$ признается возможным сценарием развития текущей ситуации.
3. Сценарии, сформированные для текущей ситуации, ранжируются по приоритетности. Наиболее приоритетный сценарий считается оптимистическим, наименее приоритетный – пессимистическим.
4. Формируются предложения по действиям, которые необходимо предпринять для способствования развитию текущей ситуации по благоприятным сценариям.

4.1 Обнаружение аналогичных событий

Событие представляет собой некоторое изменение ситуации в реальном мире. Однако текстовое описание события характеризует не только само изменение, но и его контекст, т. е. содержит информацию о ситуации в целом. Например, в сообщении о завершении тушения пожара содержится некоторая общая информация о чрезвычайной ситуации – место и время возникновения пожара, причина и условия протекания. Аналогичными будем считать события, соответствующие схожим изменениям ситуаций без учета контекста.

Для определения аналогичности события ε_i, принадлежащего ситуационному графу, и события $\varepsilon_{s_e}^j$, принадлежащего эталонной ситуации s_e, определим расстояние $\gamma_{an}(\varepsilon_i, \varepsilon_{s_e}^j)$ между ними. Функция $\gamma_{an}(\varepsilon_i, \varepsilon_j)$ принимает неотрицательные значения, причем $\gamma_{an}(\varepsilon_i, \varepsilon_j) = 0$, если события ε_i и ε_j описывают полностью идентичные изменения, произошедшие в рамках соответствующих ситуаций. Если расстояние $\gamma_{an}(\varepsilon_i, \varepsilon_{s_e}^j)$ меньше порогового значения Th_{an}, делается вывод о том, что текущее событие ε_i аналогично эталону $\varepsilon_{s_e}^j$.

При определении аналогичности событий учитываются их названия, текстовые описания и тематический состав. Текстовое описание события ε_i задается вектором $\varepsilon_i^w = (w_i^1, w_i^2, \dots, w_i^{N^w})$, где N^w – количество различных слов, встречающихся в описаниях событий, w_i^l – вес -го слова в описании i-го события, который находится методом TF-IDF.

Для того чтобы наиболее важную роль при определении аналогичности играли термы, характерные для конкретного события, а не ситуации в целом, было решено умножать вес каждого терма w_i^l в ε_i^w на коэффициент k_i^l, отражающий соотношение значимости терма для события и для ситуации s, к которой относится это событие: $k_i^l = w_i^l len(s) / \sum_{\varepsilon_j \epsilon s} w_j^l$, где $len(s)$– количество событий в ситуации s:

$$\varepsilon_i^{w\prime} = (w_i^1 k_i^1, w_i^2 k_i^2, \dots, w_i^{N^w} k_i^{N^w}).$$

Расстояние между событиями с точки зрения текста рассчитывается на основе косинусной меры: $\gamma_{i,j}^w = 1 - sim_{\cos}(\varepsilon_i^{w\prime}, \varepsilon_j^{w\prime})$. Представление слов названия события $\varepsilon_i^{tw\prime}$ и расчёт расстояния между событиями с точки зрения названий $\gamma_{i,j}^{tw}$ выполняется аналогично.

Тематический состав события характеризует вектор $\varepsilon_i^t = (t_i^1, t_i^2, \dots, t_i^{N^t})$, где N^t – количество анализируемых тем, а t_i^l – значение, отражающее релевантность -го события l-ой теме. Темы задаются экспертами в виде формализованных поисковых запросов, а значения t_i^l рассчитываются на основе модифицированного метода Okapi BM25 с помощью поисковой машины Sphinx [2]. Расстояние между событиями с точки зрения тематического состава $\gamma_{i,j}^t$ также определяется на основе косинусной меры близости векторов: $\gamma_{i,j}^t = 1 - sim_{\cos}(\varepsilon_i^t, \varepsilon_i^t)$.

Расстояние между событиями с точки зрения аналогичности может быть представлено как взвешенная сумма расстояний по различным критериям:

$$\gamma_{an}(\varepsilon_i, \varepsilon_j) = \lambda^w \gamma_{i,j}^w + \lambda^{tw} \gamma_{i,j}^{tw} + \lambda^t \gamma_{i,j}^t.$$

Нахождение значений коэффициентов λ^w, λ^{tw}, λ^t и порогового значения Th_{an} может быть выполнено путем решения задачи линейной бинарной классификации, состоящей в отнесении векторов $\gamma_{i,j} = (\gamma_{i,j}^w, \gamma_{i,j}^{tw}, \gamma_{i,j}^t)$ к одному из двух классов: один означает аналогичность сравниваемых событий, а второй – её отсутствие. Решение задачи заключается в построении разделяющей плоскости:

$$\lambda^w \gamma_{i,j}^w + \lambda^{tw} \gamma_{i,j}^{tw} + \lambda^t \gamma_{i,j}^t - Th_{an} = 0.$$

Анализируемый вектор $\gamma_{i,j}$ относится к одному из классов, исходя из его расположения относительно плоскости.

Для решения задачи может быть использована машина опорных векторов (SVM). Чтобы обеспечить возможность нахождения расстояния $\gamma_{an}(\varepsilon_i, \varepsilon_j)$ как взвешенной суммы значений $\gamma_{i,j}^w$, $\gamma_{i,j}^{tw}$ и $\gamma_{i,j}^t$, необходимо использовать SVM с линейным ядром. Для обучения машины используется набор векторов

обоих классов, подготовленный экспертами.

Описанный способ обнаружения аналогов позволяет находить для текущих событий схожие события, происходившие в прошлом. Так, для события «Власти вынудили Uber свернуть онлайн-вызов такси с автопилотом в Сан-Франциско» (Рис. 1) такими аналогами являются другие случаи запрета использования тех или иных технологий органами власти по соображениям безопасности, в частности, событие «США официально запретили продажу Samsung Galaxy Note 7». После обнаружения события-аналога выполняется попытка выделения в ситуационном графе цепочки, аналогичной соответствующей эталонной ситуации (в данном случае – ситуации, касающейся проблем Samsung, связанных со смартфоном Galaxy Note 7).

4.2 Определение близости между ситуациями

На формируемую текущую ситуацию накладывается следующее ограничение: события, аналогичные событиям из эталонной цепочки, должны следовать друг за другом в том же порядке, что и соответствующие события в эталонной ситуации. Это связано с тем, что последовательность событий в эталонной цепочке отражает их причинно-следственную связь и логику развития ситуации. Если в текущей и эталонной последовательностях события располагаются в разном порядке, значит, логика их развития различна, и они не могут быть признаны аналогами.

Таким образом, при определении близости между ситуациями необходимо учитывать, что цепочки содержат ряд попарно аналогичных событий, располагающихся в цепочках в одинаковом порядке (на Рис. 2 они выделены серым цветом, пунктирной линией соединены события-аналоги). Кроме того, каждая из ситуаций может содержать события, аналоги которых отсутствуют в другой цепочке. На Рис. 2 эталонные события, аналоги которых отсутствуют в текущей ситуации, выделены вертикальной штриховкой, «лишние» события текущей ситуации – горизонтальной. Также необходимо помнить о том, что при сравнении учитывается лишь начальная часть эталонной ситуации $st(s_e, s_c)$ – от её первого события ($\varepsilon_{s_e}^1$ на рис. 2) до последнего события, имеющего аналог в текущей ситуации ($\varepsilon_{s_e}^6$ на рис. 2). События, составляющие заключительную часть эталонной ситуации $fin(s_e, s_c)$, не влияют на значение близости.

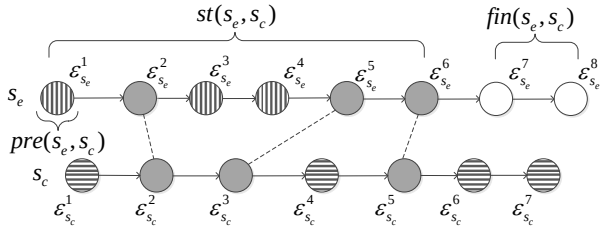


Рисунок 2 Сопоставление цепочек событий

В Таблице 1 представлен пример сравнения текущей ситуации с эталонной. В данном случае пары событий (ε_1^1,ε_2^1), (ε_1^2,ε_2^2) и (ε_1^3,ε_2^3) являются аналогами, событие ε_1^4 является «лишним» событием текущей ситуации, а событие ε_2^4 является заключительной частью эталонной ситуации.

Таблица 1 Сопоставление текущей и эталонной ситуаций

Текущая ситуация	**Эталонная ситуация**
ε_1^1: Компания Uber запустила беспилотное такси в США	ε_2^1: Выпущен Samsung Galaxy Note 7
ε_1^2: Власти Калифорнии требуют от Uber прекратить использование беспилотных такси	ε_2^2: Власти США призвали отказаться от использования Samsung Galaxy Note 7
ε_1^3: Власти вынудили Uber свернуть онлайн-вызов такси с автопилотом в Сан-Франциско	ε_2^3: США официально запретили продажу Samsung Galaxy Note 7
ε_1^4: Uber перенесла беспилотные такси в Аризону	
	ε_2^4: Samsung объявил о прекращении производства Galaxy Note 7

Для измерения близости ситуаций используется метод, представляющий собой модификацию расстояния Левенштейна: расстояние между цепочками определяется суммарным весом операций, необходимых для превращения одной цепочки в другую. Рассмотрим операции, которые требуются для превращения начальной части эталонной ситуации $st(s_e, s_c)$ в текущую ситуацию s_c, а также способы измерения веса этих операций.

- Удаление событий $\varepsilon_{s_e}^i$, аналоги которых отсутствуют в текущей ситуации. В качестве веса операции $w_{del}(\varepsilon_{s_e}^i)$ может использоваться значимость удаляемого события – показатель, учитывающий количество документов, описывающих событие, и авторитетность источников, опубликовавших эти документы. Множество удаляемых событий обозначим E_{del}. Суммарный вес таких операций: $W_{del} = \sum_{\varepsilon_{s_e} \epsilon E_{del}} w_{del}(\varepsilon_{s_e})$.
- Добавление событий $\varepsilon_{s_c}^i$, аналоги которых отсутствуют в эталонной ситуации. Вес операции $w_{add}(\varepsilon_{s_c}^i)$ вычисляется аналогично. Множество добавляемых событий обозначим E_{add}. Суммарный вес операций добавления: $W_{add} = \sum_{\varepsilon_{s_c} \epsilon E_{add}} w_{add}(\varepsilon_{s_c})$.
- Замена события из эталонной цепочки $\varepsilon_{s_e}^i$ на его аналог $\varepsilon_{s_c}^j$. Вес этой операции $w_{rep}(\varepsilon_{s_e}^i, \varepsilon_{s_c}^j)$ определяется расстоянием $\gamma_{an}(\varepsilon_{s_e}^i, \varepsilon_{s_c}^j)$ между событиями $\varepsilon_{s_e}^i$ и $\varepsilon_{s_c}^j$ с точки зрения аналогичности.

Множество пар $(\varepsilon_{s_e}^i, \varepsilon_{s_c}^j)$ обозначим P_{rep}. Суммарный вес операций этого вида: $W_{rep} = \sum_{(\varepsilon_{s_e}^i, \varepsilon_{s_c}^j)\epsilon P_{rep}} w_{rep}(\varepsilon_{s_e}^i, \varepsilon_{s_c}^j)$.

- Изменение (сокращение или удлинение) временного интервала $t_{s_e}^{i,j}$ между событиями. Интервалу $t_{s_e}^{i,j}$ в эталонной последовательности соответствует интервал $t_{s_c}^{k,l}$ в текущей ситуации, где $\varepsilon_{s_c}^k$ – аналог $\varepsilon_{s_e}^i$, а $\varepsilon_{s_c}^l$ – аналог $\varepsilon_{s_e}^j$. Вес этой операции $w_{trep}(t_{s_e}^{i,j}, t_{s_c}^{k,l})$ определяется относительной разностью между величинами интервалов: $w_{trep}(t_{s_e}^{i,j}, t_{s_c}^{k,l}) = |t_{s_e}^{i,j} - t_{s_c}^{k,l}|/t_{s_e}^{i,j}$. Множество пар $(t_{s_e}^{i,j}, t_{s_c}^{k,l})$ обозначим T_{trep}. Суммарный вес таких операций: $W_{trep} = \sum_{(t_{s_e}^{i,j}, t_{s_c}^{k,l})\epsilon T_{trep}} w_{trep}(t_{s_e}^{i,j}, t_{s_c}^{k,l})$.

Ввиду различия способов расчёта веса операций разных типов, при вычислении расстояния между цепочками они должны учитываться с различными коэффициентами. Кроме того, расстояние необходимо нормировать, поскольку, чем короче учитываемая при сравнении часть эталонной цепочки, тем меньше модифицирующих операций с ней можно произвести с сохранением аналогичности полученной последовательности оригиналу. Таким образом, расстояние между эталонной и текущей цепочкой

$$\rho(s_e, s_c) = \frac{\theta^T W}{len\big(st(s_e, s_c)\big)} = \\ = \frac{(\theta_{del} W_{del} + \theta_{add} W_{add} + \theta_{rep} W_{rep} + \theta_{trep} W_{trep})}{len(st(s_e, s_c))},$$

где $len(st(s_e, s_c))$ – количество событий в начальной части $st(s_e, s_c)$ эталонной ситуации s_e, а θ_{del}, θ_{add}, θ_{rep} и θ_{trep} – коэффициенты, определяющие вклад операций различных типов в значение расстояния.

4.3 Определение вероятности аналогичности ситуаций

На основе расстояния $\rho(s_e, s_c)$ необходимо определить, является ли текущая ситуация аналогом эталонной и какова вероятность того, что текущая ситуация будет развиваться по сценарию, определяемому эталонной ситуацией. С этой целью было принято решение рассмотреть сравнение цепочек как задачу логистической регрессии. Для этого введем переменную y, принимающую одно из двух возможных значений:

$$y = \begin{cases} 1, \text{если цепочки не являются аналогами,} \\ 0, \text{если цепочки являются аналогами.} \end{cases}$$

Предположим, что вероятность наступления события $y = 0$ (т. е. вероятность того, что текущая ситуация является аналогом эталонной) задана функцией:

$$\mathrm{P}(y = 0|s_e, s_c) = 1 - \frac{1}{1 + \exp\left(-\frac{\theta^T W}{len\big(st(s_e, s_c)\big)}\right)}.$$

Значения параметров θ подбираем методом максимального правдоподобия на основе обучающей выборки, состоящей из множества пар аналогичных и неаналогичных ситуаций.

Логистическая регрессия позволяет также выполнить бинарную классификацию пар ситуаций: цепочки s_e и s_c считаются потенциальными аналогами при $\mathrm{P}(y = 0|s_e, s_c) > 0.5$.

4.4 Формирование сценария

Построение текущей ситуации начинается с нового или измененного события ситуационного графа ε_c, которое обязательно должно ей принадлежать. Далее на каждом шаге выполняется попытка дополнить ситуацию путем присоединения к цепочке одного из соседей события, которое на данный момент является первым или последним в цепочке. При этом необходимо рассмотреть различные варианты интерпретации добавляемого в цепочку события. Оно может интерпретироваться и как аналог некоторого события из s_e, и как «лишнее» событие, не имеющее аналогов в эталонной цепочке. Путем выбора на каждом шаге одного из возможных событий, добавляемых в цепочку, а также одного из возможных вариантов его интерпретации формируется дерево возможных вариантов построения текущей ситуации. Из всех вариантов текущей ситуации, рассмотренных в процессе построения, выбирается цепочка s_c^{max}, имеющая максимальную близость к эталону. Эта последовательность считается завершенной текущей ситуацией.

Если $\mathrm{P}(y = 0|s_e, s_c^{max}) > 0.5$, полученная текущая ситуация признается аналогом s_e. В этом случае $fin(s_e, s_c^{max})$ считается возможным сценарием дальнейшего развития текущей ситуации, а значение $\mathrm{P}(y = 0|s_e, s_c^{max})$ – вероятностью того, что текущая ситуация будет развиваться в соответствии с этим сценарием. На основе всех эталонных ситуаций, аналогичных текущей, формируется множество возможных сценариев её дальнейшего развития. Заключительная часть цепочки, для которой вероятность аналогичности текущей ситуации максимальна ($s_e^{prob} = \underset{s_e}{\mathrm{argmax}}[\mathrm{P}(y = 0|s_e, s_c^{max})]$), является наиболее вероятным сценарием.

4.5 Выделение оптимистического и пессимистического сценариев

Для выделения оптимистического и пессимистического сценариев необходимо определить оптимальность каждого из них. Для этого используется метод анализа иерархий (МАИ), позволяющий определить приоритет различных альтернатив с точки зрения цели с учетом различных критериев [17]. Целью в данном случае является выбор оптимального сценария, альтернативами – сформированные сценарии, а в качестве критериев могут использоваться такие характеристики сценариев, как длительность, экономическая эффективность и другие. Выбор критериев определяется предметной областью, в рамках которой используется прогнозирование развития ситуаций.

Значения критериев для эталонных ситуаций определяются экспертами на этапе подготовки базы эталонов S_e. Также эксперты путем попарных сравнений определяют приоритетность критериев относительно цели. Приоритетность сценариев относительно каждого из критериев может быть определена автоматически при анализе ситуационного графа на основе сравнения характеристик соответствующих эталонных ситуаций. Это позволяет автоматически определить приоритет относительно цели для каждого из сценариев, сформированных для текущей ситуации. Сценарий с максимальным приоритетом считается оптимистическим, сценарий с минимальным приоритетом – пессимистическим.

На Рис. 3 показаны оптимистический и пессимистический сценарии, сформированные для ситуации, связанной с тестированием беспилотных такси. Для определения приоритетности сценариев использовались такие критерии, как «безопасность», «длительность» и «экономическая эффективность». Также на рисунке представлен наиболее вероятный сценарий, определенный с помощью логистической регрессии.

4.6 Формирование предложений для лиц, принимающих решения

С целью последующего формирования предложений эксперты должны снабжать каждое событие ε_e каждой эталонной ситуации s_e рекомендациями по действиям, которые должны предприниматься при наступлении аналогичного события в будущем. Рекомендация $rec_{\varepsilon_e} = \langle action_{\varepsilon_e}, actor_{\varepsilon_e}, period_{\varepsilon_e} \rangle$ содержит информацию о действиях $action_{\varepsilon_e}$, которые должны быть предприняты лицом $actor_{\varepsilon_e}$ в срок $period_{\varepsilon_e}$ для содействия или противодействия развитию текущей ситуации по сценарию, сформированному на основе s_e.

На Рис. 3 показаны рекомендации для ЛПР с учетом сценариев, сформированных для ситуации с тестированием беспилотных такси.

Рисунок 3 Пример формирования оптимистического, наиболее вероятного и пессимистического сценариев развития ситуации

5 Экспериментальная проверка метода

На основе предложенного метода разработана система автоматизированного мониторинга и прогнозирования развития ситуаций. Обучение системы выполняется экспертами на основе эталонных событий и ситуаций. Обученная система автоматически обрабатывает текстовый поток, обнаруживает события и формирует ситуации, а также определяет вероятные сценарии их дальнейшего развития и вырабатывает рекомендации.

Результаты качества работы подсистемы обнаружения событий приведены в [5]. Эксперименты показали, что при использовании для обучения 1300 пар документов и событий достигается значение точности 85,2%, полноты – 76% и F-меры – 79,8%.

Для анализа качества работы подсистемы формирования сценариев был проведен эксперимент с целью определения зависимости точности, полноты и F-меры выявления аналогичных ситуаций от мощности обучающей выборки. Полученные зависимости приведены на Рис. 4. В результате проведения эксперимента оказалось, что для обучения системы достаточно 90 пар ситуаций. При таком количестве обучающих примеров достигается значение F-меры около 0.8, с дальнейшим увеличением обучающей выборки качество работы метода не улучшается.

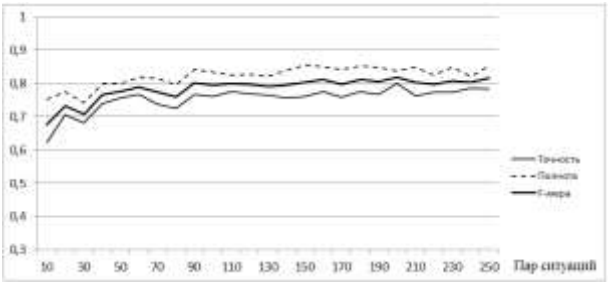


Рисунок 4 Зависимость точности (тонкая сплошная линия), полноты (пунктирная линия) и F-меры (жирная линия) от мощности обучающей выборки

6 Направления дальнейших исследований

Предложенный метод прогнозирования развития ситуаций предоставляет пользователю сценарии дальнейшего развития ситуации и рекомендации по действиям, необходимым для их реализации, но не позволяет осуществлять управление развитием ситуации по оптимальному сценарию. Пользователю требуется определять, соответствует ли развитие ситуации сформированному ранее сценарию, и получать рекомендации в случае необходимости корректировки намеченного плана мероприятий. В связи с этим дальнейшим направлением развития метода является разработка более сложных сетевых моделей эталонных ситуаций, способных отражать различные варианты возможного развития текущей ситуации в зависимости от действий ЛПР на каждом этапе управления ситуацией.

Выше описан эксперимент по оценке качества обнаружения аналогичных ситуаций, однако необходимо также оценивать качество прогнозирования. В связи с этим в рамках дальнейших исследований планируется выработать критерий качества ситуационного прогноза и выполнить оценку результатов прогноза по этому критерию.

7 Заключение

Предложен метод прогнозирования развития ситуаций на основе обнаружения событий в потоке текстовых документов. Прогнозирование состоит в формировании сценариев дальнейшего развития ситуаций по принципу исторической аналогии: выполняется построение текущей ситуации, для которой существует аналог в базе эталонных ситуаций. Этот аналог считается возможным сценарием развития текущей ситуации. Предложенный метод формирования сценариев учитывает динамику развития ситуаций и нестрогий характер аналогии между ситуациями. Из множества сформированных сценариев выделены оптимистический и пессимистический, для этого использован метод анализа иерархий. Также предложен способ подготовки предложений по действиям, которые необходимо предпринять для способствования или препятствования развитию ситуации по построенным сценариям.

Литература

[1] Aggarwal, C.C., Subbian, K.: Event Detection in Social Streams. Proc. of the 2012 SIAM Int. Conf. on Data Mining, pp. 624-635. Society for Industrial and Applied Mathematics, Philadelphia (2012). doi: 10.1137/1.9781611972825.54

[2] How Sphinx Relevance Ranking Works. http://sphinxsearch.com/blog/2010/08/17/how-sphinx-relevance-ranking-works/

[3] van der Aalst, W.M.P.: Process Mining: Data Science in Action. Springer, Heidelberg (2016). doi: 10.1007/978-3-662-49851-4

[4] Андреев, А.М., Березкин, Д.В., Брик, А.В., Смирнов, Ю.М.: Вероятностный синтаксический анализатор для информационно-поисковых систем. Вестник МГТУ. Сер. Приборостроение, 2, сс. 34-53 (2000)

[5] Андреев, А.М., Березкин, Д.В., Козлов, И.А.: Подход к автоматизированному мониторингу тем на основе обнаружения событий в потоке текстовых документов. Информационно-измерительные и управляющие системы, 15 (3), сс. 49-60 (2017)

[6] Андреев, А.М., Березкин, Д.В., Симаков, К.В.: Обучение морфологического анализатора на большой электронной коллекции текстовых документов. Электронные библиотеки: перспективные методы и технологии, электронные коллекции: Труды Седьмой Всерос. науч. конф. (RCDL–2005), сс. 173-181 (2005)

[7] Андреев, А.М., Березкин, Д.В., Симаков, К.В.: Особенности проектирования модели и онтологии предметной области для поиска противоречий в правовых электронных библиотеках. Электронные библиотеки: перспективные методы и технологии, электронные коллекции: Труды Шестой Всерос. науч. конф. RCDL, сс. 93-102 (2004)

[8] Ахременко, А.С.: Политический анализ и прогнозирование: Учеб. пособие. М.: Гардарики (2006)

[9] Борисов, В.В., Зернов, М.М.: Реализация ситуационного подхода на основе нечеткой иерархической ситуационно-событийной сети. Искусственный интеллект и принятие решений, 1, сс. 18-30 (2009)

[10] Варшавский, П.Р.: Методы и программные средства поиска решения на основе аналогий в интеллектуальных системах поддержки принятия решения. Дисс. … канд. техн. наук, Московский энергетический институт (2005)

[11] Волгин, Н.С.: Исследование операций, ч. 1. С-Пб.: ВМА им. Н. Г. Кузнецова (1999)

[12] Еремеев, А.П., Варшавский, П.Р.: Моделирование рассуждений на основе прецедентов в интеллектуальных системах поддержки принятия решений. Искусственный интеллект и принятие решений, 2, сс. 45-57 (2009)

[13] Зацаринный, А.А., Сучков, А.П.: Некоторые подходы к ситуационному анализу потоков событий. Открытое образование, 1, сс. 39-46 (2012)

[14] Кононов, Д.А., Косяченко, С.А., Кульба, В.В.: Формирование и анализ сценариев развития социально-экономических систем с использованием аппарата операторных графов. Автоматика и телемеханика, 68 (1), сс. 121-136 (2007)

[15] Косяченко, С.А., и др.: Модели, методы и автоматизация управления в условиях чрезвычайных ситуаций. Автоматика и телемеханика, 59 (6), сс. 3-66 (1998)

[16] Кулинич, А.А.: Компьютерные системы моделирования когнитивных карт: подходы и методы. Проблемы управления, 3, сс. 2-16 (2010)

[17] Саати, Т.: Методы анализа иерархий. М.: Радио и связь (1993)

[18] Ситчихин, А.Н.: Иерархические ситуационные модели с предысторией для автоматизированной поддержки решений в сложных системах. Дисс. … канд. техн. наук, Уфимский гос. авиационный технический университет (2002)

Применение машинного обучения

Application of machine learning

Machine Learning Methods Application to Search for Regularities in Chemical Data

© N.N.Kiselyova[1] ©A.V.Stolyarenko[1] ©V.A.Dudarev[1,2]

[1]Institution of Russian Academy of Sciences A.A. Baikov Institute of Metallurgy and Materials Science RAS (IMET RAS),
Moscow, Russia
[2]National Research University Higher School of Economics (NRU HSE),
Moscow, Russia

kis@imet.ac.ru stol-drew@yandex.ru vic@imet.ac.ru

Abstract. The possibility of searching for classification regularities in large arrays of chemical information by means of machine learning methods is discussed. Tasks peculiarities in inorganic chemistry and materials science are considered. The short review of these methods applications to inorganic chemistry and materials science is presented. The system for computer-assisted inorganic compounds design based on machine learning methods has been developed. The developed system usage makes it possible to predict new inorganic compounds and estimate some of their properties without experimental synthesis. The results of this information-analytical system application to inorganic compounds design are promising for new materials search.

Keywords: machine learning, database, inorganic chemistry, design of inorganic compounds.

1 Introduction

Throughout the centuries of its evolution chemistry and materials science accumulated huge information. In common with other experimental sciences chemistry got through several stages: information accumulation, data analysis and development of classification schemes and rules that allow classifying a new object to a particular substances class. The substances division into inorganic and organic ones, Periodic table of elements, compounds classification according to crystal structure type, etc. are examples of such classifications. Essentially, in all cases these classifications are imprecise, and classes intersect partially. For example, organic chemistry is determined as the carbon compounds chemistry but carbides and carbonates belong to inorganic chemistry objects as well as boron hydrides (boranes) or silicon hydrides (silanes) which are closer to hydrocarbons (organic chemistry objects) in many properties. In large measure, it is caused by imperfection in the classification rules which were developed by chemists. One way to get around these problems in inorganic chemistry and materials science is machine learning methods application to information analysis aimed at discovery of complicated classifying regularities that allow considering of substances to particular classes. It is noteworthy that obtained regularities include substance components properties as variables, and for this reason, their use allows us to predict the class for the substances that is not yet synthesized knowing only the well-known parameters values for chemical elements forming this substance.

Half a century ago IMET pioneered in applying such approach to machine learning use to search for classifying regularities that allows a prediction of new inorganic compounds and some of their properties estimation [1, 2] to be made. The machine learning methods application [3, 4] made possible allowed new binary compounds prediction with 90% reliability knowing constituent chemical elements properties only. The success of approach that was put forward in IMET has given an impetus to many investigations which were connected with machine learning application to inorganic chemistry and materials science and carried-out in various countries. The investigations geography in this field is very wide: Europe, America, Asia, Africa (figure 1). The most representative teams work in Russia, the USA, and China. More detailed reviews of these researches are given in the monograph [5] and reviews [6, 7]. It should be noted that in recent years in the developed countries the governmental initiatives aimed at IT application (as well as machine learning methods) to chemistry and materials science were announced: Materials Genome Initiative (the USA) [8], Materials Research by Information Integration Initiative (Japan) [9], and Chinese Materials Genome (China) [10]. It is expected that the theoretic methods use will provide essential progress achievements in chemistry and materials science that will lead to cost reduction during new materials research, development, and production.

2 Problem Statement and Decision Methods

Suppose that every inorganic substance is described by a vector $\mathbf{x} = (x_1^{(1)}, x_2^{(1)}, \ldots x_M^{(1)}, x_1^{(2)},$

Proceedings of the XIX International Conference "Data Analytics and Management in Data Intensive Domains" (DAMDID/RCDL'2017), Moscow, Russia, October 10–13, 2017

$x_2^{(2)}$,… $x_M^{(2)}$,…, $x_1^{(L)}$, $x_2^{(L)}$,... $x_M^{(L)}$), where L is the number of chemical elements that form a compound and M is the number of chemical elements parameters. Each substance is also characterized by a class membership parameter: a(x) ∈ {1, 2,…, K}, where K is the number of classes. The learning sample consists of N objects: S = {x_i, i = 1, …, N}. We denote the learning sample objects subset from class a_j, j = 1, 2, …, K, by S_{aj} = {x: a(x) = a_j}. The machine learning aim is to construct a classification rule that distinguishes not only different classes objects of the learning sample but also preserves prognostic ability to generate new combinations of chemical elements that were not used for learning.

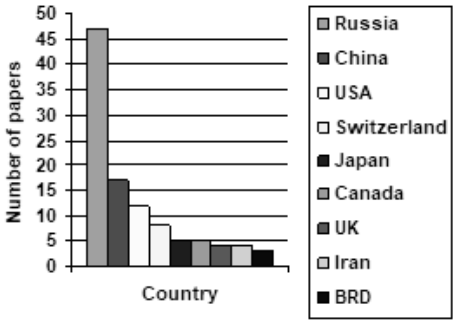


Figure 1 Distribution of publications related to machine learning methods applications to inorganic chemistry and materials science over the countries

Among the numerous machine learning methods, various of Artificial Neural Network (ANN) learning algorithms modifications and Support Vector Machine algorithms (SVM) are the most popular (figure 2). This is due to appropriate software packages accessibility and seeming exam score accuracy (many investigators do not take into account an influence of overfitting effect on subsequent prediction reliability that is inherent in these methods).

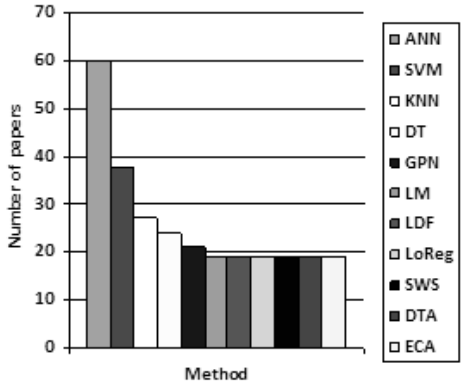


Figure 2 Various machine learning methods popularity in inorganic chemistry

Notation: ANN – artificial neural network learning; SVM – support vector machine; KNN – k-nearest neighbors method; DT – decision trees learning; GPN – concept formation using growing pyramidal networks; LM – linear machine method; LDF – linear Fisher discriminant; LoReg – voting algorithm where estimations for classes are calculated by means of voting by logical regularities system; SWS – statistical weighted syndromes; DTA – deadlock test algorithm; ECA – estimate calculating algorithm.

A great diversity of chemical and materials science tasks were solved successfully using machine learning methods, e.g.:

theoretic tasks of prediction of:

- inorganic system phase diagram type [5, 11];
- inorganic compounds formation with certain stoichiometric composition [1, 2, 5-7, 12];
- inorganic compounds crystal structure type [5-7, 13, 14];
- some of inorganic compounds properties (melting point [15], critical temperature of superconductivity [16], band gap energy [17], enthalpy of formation [18], etc.);

technologic tasks of prediction of:

- mechanical properties of steels [19];
- acoustic properties of tellurite glasses [20];
- tribological behavior of aluminum–copper based composite [21];
- functional properties of ceramic materials [22], and so on.

3 Experience in machine learning system development for chemical applications

A special information-analytical system (IAS) that allows an automation of task solution procedure in the field of inorganic chemistry using machine learning was developed in IMET [23]. The subject field peculiarities were taken into account at the IAS creation, namely:

1) Attribute description composite structure: chemical elements (inorganic substance components) parameters set is repeated as many times as the number of elements which are included into the compound.

2) Strong correlation within set of these attributes for each component due to their dependence on common parameter - chemical elements atomic number (it follows from the Periodic Law).

3) Individual chemical elements properties give small informative gain therefore more informative parameters of single compounds (for example, single oxides, halogenides, chalcogenides, etc.) and component properties algebraic functions are widely used additionally.

4) Blanks of attributes' values that are filled by various methods including interpolation taking into account a periodicity in chemical elements properties variation with their atomic numbers.

5) Large asymmetry of learning set sizes for different classes (at that often the least of representative – as a rule newly obtained classes of substances – are the most interesting for chemists).

6) Errors and discrepancies in inorganic compounds experimental classification of learning set decreases the prediction accuracy drastically.

Machine learning procedure involves several stages:

1) objects selection for machine learning,

2) attribute description formation (including the most informative attributes selection and filling attribute values blanks also),

3) machine learning algorithms selection,

4) machine learning including application of algorithms ensembles and collective solution synthesis in a case of several algorithms usage,

5) machine learning quality estimation,

6) new objects status prediction and results interpretation.

3.1 Objects selection for machine learning

Representative and reliable set formation for machine learning preconditions subsequent prediction accuracy in a great measure. Objects selection (known inorganic substances examples) for machine learning is performed by experts in subject domain by means of information stored in data bases (DBs) on inorganic substances and materials properties including DBs that were developed in IMET [17, 23-25]. The latest include data on tens of thousands of substances and are Internet-accessible [25]. Data on substances were extracted from thousands of publications. In common with other intellectual fields papers can involve errors and inaccuracies. The experimental errors in object classification contribute significantly to prediction accuracy decrease. However, classification reliability estimation of tens of thousands of substances is massively expensive and practically impossible task. Partial automation of procedure of search for data outliers using machine learning is proposed by us. This can be best done in detecting of errors which were caused by incorrect and incomplete experimental knowledge of the class to which the substance belongs (for example, crystal structure type) as well as by erroneous property values of components which form the substance description. In the latter case errors can be incorrect experimental property value measurement result or they can be associated with incorrect interpolation in the case of filling attribute values blanks as well. The machine learning results analysis allows detection of substances which fall within another class and provision for chemist with information on substance expert assessment and making a decision for its status. The problem solution principal possibility is specified by the subject domain specific that is connected with inorganic compounds properties variation periodicity depending on atomic number of elements – the chemical system components.

3.2 Attribute description formation

Attribute description formation problem is complicated and hard-to-solve task of modern machine learning theory. There are a large number of approaches which have proved their effectiveness at various task types solution. However, it is impossible to evolve a surely optimal universal method of attributes selection. In this regard few alternative methods with subsequent collective decision synthesis are used by us for attribute selection. 2D-projections visualization tools are applied additionally for points corresponding to certain type compounds in chemical elements properties space. The parameters set includes not only initial attributes but also the algebraic functions of these attributes which are selected by user.

3.3 Machine learning algorithms selection

The IAS includes a set of machine learning algorithms which are the most popular among chemists (figure 1). At present time IAS involves the following software: programs based on well-known linear machines methods, Fisher linear discriminant, k-nearest neighbors, support vector machine, neural-network algorithms, and also algorithms which were developed by the Computing Centre, Russian Academy of Sciences and based on estimates calculation, deadlock tests voting algorithms, logical regularities voting algorithms, weighted statistical voting algorithms, etc. [26]. IAS includes also the ConFor system for machine learning according to procedure for concept formation, developed by the Institute of Cybernetics, National Academy of Sciences of Ukraine [27]. This system is built upon computer memory data arrangement in the form of growing pyramidal networks. At solution of each task at hand a selection of the most exact machine learning algorithms is carried out for subsequent use in decision making and prediction procedures.

3.4 Machine learning

Our experience in inorganic chemistry prediction tasks solution shows [6, 7, 12, 17, 23, 24] that algorithms ensembles application allows a considerable increase of accuracy in inorganic compounds prediction. In decision making process the most accurate machine learning algorithms are used that were selected on the previous stage. The IAS includes the following programs realizing various collective decisions strategies, which are based on Bayes method, clustering and selection methods, decision templates, logical correction, convex stabilizer method, Woods dynamic method, committee methods, etc. [26].

3.5 Machine learning quality estimation

The cross-validation on learning set objects is the most widely used universal and reliable tool for machine learning quality estimation. IAS contains special software for this procedure realization that is used in the best machine learning algorithms selection. However, an attempt of cross-validation application to machine learning accuracy estimation at use of algorithms ensembles as optimizable criterion results the loss of estimate unbiasedness. In this case, there is a certain overfitting risk. In this regard, the traditional approach to collective algorithms accuracy evaluation using examination recognition of N examples chosen randomly from learning samples and unused in learning (at the final prediction stage, reference examples are returned to the learning set) is applied. The corresponding program was included to IAS.

The learning set sizes asymmetry for different classes is an important problem at machine learning accuracy estimation. Naturally in this case the generalized examination recognition accuracy does not represent the prediction error for small classes, therefore the ROC curves application is appropriate to different algorithms prediction quality analysis. ROC curves allow recognition accuracy comparison for the targeted and alternative classes at variation of cut-offs which identifies belonging to different classes.

It should be pointed out that machine learning quality estimation procedure belongs to yet hardly unsolved machine learning task. Some algorithms (SVM, ANN, etc.) characterized by overfitting effect, show high examination recognition accuracy often but this fact does not always provide high predicting reliability for new objects.

3.6 Prediction of new inorganic compounds formation and some of their properties estimation

To increase predicting accuracy in the case of learning sets with K classes (K > 2) the following method is used. Firstly, multi-class learning and prediction are carried out. Next, K dichotomies are calculated: the targeted class and all the alternative classes, followed by subsequent K predictions. The results of multi-class prediction and dichotomies series are intercompared, and if the predictions are not contradictory the decision on the object status is made. The special tools for collective decision formation based on comparison of multi-class prediction results and dichotomies series were developed. The efficiency of such approach that allows to increase prediction accuracy was approved during numerous tasks solution [5-7, 12, 17, 23, 24].

4 IAS application illustration to regularities search in chemical information

The machine learning application allowed a search for inorganic compounds formation regularities, a prediction of thousands not yet synthesized substances and some their properties estimation using obtained regularities. This approach efficiency to inorganic compounds design can be illustrated by comparison of the predictions results with newer experimental data obtained after publication of our predictions [12].

The table contains AB_3X_3 compounds formation possibility predictions in the A_2X_3–B_2X systems (A and B are various elements, and X = S, Se, or Te) under normal conditions, which could be promising for search for new semiconductor, nonlinear optical, electro-optical, and acousto-optical materials. Experimental information on 117 examples of AB_3X_3 compounds formed and 58 examples when no such composition compounds were formed in the A_2X_3–B_2X systems under normal conditions was used for computer analysis. To describe the compounds in computer memory we selected A, B, and X elements properties (the melting and boiling points; covalent, ionic (by Bokii and Belov), and pseudopotential (by Zunger) radii; the first three ionization potentials; electronegativity (by Pauling); the standard enthalpies of atomization and evaporation; thermal conductivity; molar heat capacities, etc.), simple A_2X_3 and B_2X chalcogenides properties (standard entropy and enthalpy), and some algebraic functions of these properties (for example, the ratio of the covalent radius to the metal radius for elements A, B, and X). The table 1 presents predictions examples for the AB_3X_3 compounds and their experimental verification results. The following notation is used: 1, the prediction of AB_3X_3 formation under normal conditions; 2, the prediction of AB_3X_3 absence under normal conditions; #, examples, the information on which is used for machine learning; empty cells, uncertain prediction; ©, the prediction of AB_3X_3 formation matches new experimental data; and Θ, the prediction of compound absence matches experimental data. All 27 tested predictions coincided with the experimental data.

Table 1 AB_3X_3 compounds formation possibility prediction

A / B	Fe	Ga	In	Sn	Sb	La	Ce	Pr	Nd	Sm	Eu	Gd	Tb	Dy	Bi
X = S															
K	©	#2	1	1	#1	1		1	1	1	1	1	1	1	#2
Rb	©		#1	1	©	1			1	1	1	1	1	1	#1
Tl	1	Θ	#1	#1	#1	#2	2	#2	Θ	2	2				
X = Se															
K	#1	#1	1	©	#1		1			1	1	1	1	1	#1
Rb	1	1	1	1	©	1	1			1	1	1	1	1	#1
Ag	2	#2	Θ		#2	Θ	Θ	Θ	2	Θ	Θ		Θ	2	Θ
Cs	1	#1	1	1	©	1				1	1	1	1	1	#1
Tl	1	Θ	#2	#1	#1	2	2	2	2						#2
X = Te															

Rb	1	1	1	©	1	1	1			1	1	1	1	1	1
Ag	2	#2	#2	Θ	#2	2	2	2	2	2	2	#2	Θ	Θ	Θ
Cs	1	1	1	©	1	1	1			1	1	1	1	1	1
Tl	©	Θ	Θ	#1	#2	#2	2	2	Θ						#2

Conclusions

During half of the century the predictions of thousands of inorganic compounds in binary, ternary and more complicated chemical systems were obtained and some their properties (melting point, critical temperature of superconductivity, band gap energy, etc.) were estimated in IMET [1, 2, 5-7, 12, 16, 17, 23, 24]. The obtained predictions usage allows an essential progress provision in a search for new magnetic, semiconductor, superconductor, nonlinear optical, electro-optical, acousto-optical and other materials. Hundreds of predicted compounds were synthesized and our results experimental verification shows that the average prediction accuracy is higher than 80% [2, 5-7]. Machine learning methods application to search for regularities in big chemical data gives an opportunity for theoretic design of new inorganic compounds that allows substantially reduce the costs for search for new materials with predefined properties, replacing them by computations. It is important to note that only information on components properties (chemical elements or more simple compounds) is used in prediction process.

This work was partially supported by the Russian Foundation for Basic Research (project nos. 16-07-01028, 17-07-01362, and 15-07-00980). We are grateful to V.V. Ryazanov, O.V. Sen'ko and A.A. Dokukin for long-term help and collaboration.

References

[1] E. M. Savitskii, Yu. V. Devingtal', and V. B. Gribulya. Prediction of metallic compounds with composition A3B using computer. Dokl. Akad. Nauk SSSR (English translation - Doklady Physical Chemistry), 183(5), p.1110-1112, 1968

[2] E. M. Savitskii and V. B. Gribulya. Application of computer techniques in the prediction of inorganic compounds. New Delhi-Calcutta: Oxonian Press Pvt., Ltd. 1985

[3] Yu. V. Devingtal'. About optimal coding of objects at their classification using pattern recognition methods. Izvestiya Akademii Nauk SSSR. Tekhnicheskaya Kibernetika, 1, p.162-169, 1968.

[4] Yu. V. Devingtal'. Coding of objects at application of separating hyper-plane for their classification. Izvestiya Akademii Nauk SSSR. Tekhnicheskaya Kibernetika, 3, p.139-147, 1971

[5] N.N. Kiselyova. Komp'yuternoe konstruirovanie neorganicheskikh soedinenii. Ispol'zovanie baz dannykh i metodov iskusstvennogo intellekta (Computer Design of Inorganic Compounds: Use of Databases and Artificial Intelligence Methods). Moscow: Nauka, 2005

[6] G.S. Burkhanov and N.N. Kiselyova. Prediction of intermetallic compounds, Russ. Chem. Rev., 78(6), p. 569-587, 2009

[7] N. Kiselyova, A. Stolyarenko, V. Ryazanov, et al. Application of Machine Training Methods to Design of New Inorganic Compounds. In Diagnostic Test Approaches to Machine Learning and Commonsense Reasoning Systems. Ed. By X.A. Naidenova & D.I. Ignatov. Hershey: IGI Global, p. 197-220, 2012

[8] Site of Materials Genome Initiative: https://www.mgi.gov/

[9] Site of Center for Materials Research by Information Integration: http://www.nims.go.jp/eng/research/MII-I/index.html

[10] X.-G. Lu. Remarks on the recent progress of Materials Genome Initiative, Sci. Bull., 60(22), p.1966–1968, 2015

[11] P. Villars, K. Brandenburg, M. Berndt, et al. Binary, ternary and quaternary compound former/nonformer prediction via Mendeleev number, J. Alloys and Compounds, 317-318, p.26-38, 2001

[12] N.N. Kiselyova. Prediction of Formation of AB3X3 (X = S, Se, Te), Inorg. Mater., 45(10), p.1077-1080, 2009

[13] A.O. Oliynyk, E. Antono, T.D. Sparks et al. High-Throughput Machine-Learning-Driven Synthesis of Full-Heusler Compounds, Chem. Mater., 28(20), p.7324−7331, 2016

[14] G. Pilania, P.V. Balachandran, J.E. Gubernatis, and T. Lookman. Classification of ABO3 perovskite solids: a machine learning study, Acta Crystallogr., B71(5), p.507-513, 2015

[15] A. Seko, T. Maekawa, K. Tsuda, and I. Tanaka. Machine learning with systematic density-functional theory calculations: Application to melting temperatures of single- and binary-component solids, Phys. Rev., B89(5), p.054303/1-9, 2014

[16] E.M. Savitskii, V.B. Gribulya, and N.N. Kiselyova. Cybernetic prediction of superconducting compounds, CALPHAD, 3(3), p.171-173, 1979

[17] N.N. Kiselyova, V.A. Dudarev, M.A. Korzhuyev. Database on the Bandgap of Inorganic Substances and Materials, Inorganic Materials: Applied Research, 7(1), p. 34-39, 2016

[18] S.P. Sun, D.Q. Yi, Y. Jiang, et al. Prediction of formation enthalpies for Al2X-type intermetallics using back-propagation neural network, Mater. Chem. and Phys., 126(3), p. 632–641, 2011

[19] A. Bahrami, A. S. H. Mousavi, and A. Ekrami. Prediction of mechanical properties of DP steels using neural network model, J. Alloys and Compounds, 392(1-2), p.177-182, 2005

[20] M.S. Gaafar, M.A.M. Abdeen, and S.Y. Marzouk. Structural investigation and simulation of acoustic properties of some tellurite glasses using artificial intelligence technique, J. Alloys and Compounds, 509, p. 3566-3575, 2011

[21] M. Hayajneh, A.M. Hassan, A. Alrashdan, and A.T. Mayyas. Prediction of tribological behavior of aluminum–copper based composite using artificial neural network, J. Alloys and Compounds, 470, p. 584-588, 2009

[22] D.J. Scott, P.V. Coveney, J.A. Kilner, et al. Prediction of the functional properties of ceramic materials from composition using artificial neural networks, J. Eur. Ceram. Soc., 27(16), p. 4425–4435, 2007

[23] N.N. Kiselyova, A.V. Stolyarenko, V.V. Ryazanov, et al. A system for computer-assisted design of inorganic compounds based on computer training, Pattern Recognition and Image Analysis, 21(1), p. 88-94, 2011

[24] N.N. Kiselyova, V.A. Dudarev, and V.S.Zemskov. Computer information resources in inorganic chemistry and materials science, Russ. Chem. Rev., 79(2), p. 145-166, 2010

[25] Site of IMET RAS DBs: http://imet-db.ru

[26] Yu. I. Zhuravlev, V. V. Ryazanov, and O. V. Sen'ko. RECOGNITION. Mathematical methods. Software system. Practical solutions. Moscow: Phasis. 2006

[27] V.P. Gladun. Processes of formation of new knowledge. Sofia: SD "Pedagog 6". 1995

Astrophysical Data Analytics based on Neural Gas Models, using the Classification of Globular Clusters as Playground

© Giuseppe Angora[1] © Massimo Brescia[2] © Giuseppe Riccio[2] © Stefano Cavuoti[3]
© Maurizio Paolillo[3] © Thomas H. Puzia[4]

[1] Department of Physics "E. Pancini", University Federico II,
Via Cinthia 6, 80126 Napoli, Italy
[2] INAF Astronomical Observatory of Capodimonte,
Via Moiariello 16, 80131 Napoli, Italy
[3] Department of Physics "E. Pancini", University Federico II,
Via Cinthia 6, 80126 Napoli, Italy
[4] Institute of Astrophysics, Pontificia Universidad Católica de Chile,
Av. Vicuña Mackenna 4860, Macul, Santiago, Chile

gius.angora@gmail.com

Abstract. In Astrophysics, the identification of candidate Globular Clusters through deep, wide-field, single band HST images, is a typical data analytics problem, where methods based on Machine Learning have revealed a high efficiency and reliability, demonstrating the capability to improve the traditional approaches. Here we experimented some variants of the known Neural Gas model, exploring both supervised and unsupervised paradigms of Machine Learning, on the classification of Globular Clusters, extracted from the NGC1399 HST data. Main focus of this work was to use a well-tested playground to scientifically validate such kind of models for further extended experiments in astrophysics and using other standard Machine Learning methods (for instance Random Forest and Multi Layer Perceptron neural network) for a comparison of performances in terms of purity and completeness.

Keywords: data analytics, astroinformatics, globular clusters, machine learning, neural gas.

1 Introduction

The current and incoming astronomical synoptic surveys require efficient and automatic data analytics solutions to cope with the explosion of scientific data amounts to be processed and analyzed. This scenario, quite similar to other scientific and social contexts, pushed all communities involved in data-driven disciplines to explore data mining techniques and methodologies, most of which connected to the Machine Learning (hereafter ML) paradigms, i. e. supervised/unsupervised self-adaptive learning and parameter space optimization[3],[6],[7] .

Following this premise, this paper is focused on the investigation about the use of a particular kind of ML methods, known as Neural Gas (NG) models[21], to solve classification problems within the astrophysical context, characterized by a complex multi-dimensional parameter space. In order to scientifically validate such models, we decided to approach a typical astrophysical playground, already solved with ML methods [8], [11] and to use in parallel other two ML techniques, chosen among the most standard, respectively, Random Forest [5] and Multi Layer Perceptron Neural Network[23], as comparison baseline.

Proceedings of the XIX International Conference "Data Analytics and Management in Data Intensive Domains" (DAMDID/RCDL'2017), Moscow, Russia, October 10–13, 2017

The astrophysical case is related to the identification of Globular Clusters (GCs) in the galaxy NGC1399 using single band photometric data obtained through observations with the Hubble Space Telescope (HST) [8], [25],[27].

The physical identification and characterization of a Globular Cluster (GC) in external galaxies is considered important for a variety of astrophysical problems, from the dynamical evolution of binary systems, to the analysis of star clusters, galaxies and cosmological phenomena [27].

Here, the capability of ML methods to learn and recognize peculiar classes of objects, in a complex and noising parameter space and by learning the hidden correlation among object's parameters, has been demonstrated particularly suitable in the problem of GC classification[8]. In fact, multi-band wide-field photometric data (colours and luminosities) are usually required to recognize GCs within external galaxies, due to the high risk of contamination of background galaxies, which appear indistinguishable from galaxies located few Mpc away, when observed by ground-based instruments. Furthermore, in order to minimize the contamination, high-resolution space-borne data are also required, since they are able to provide particular physical and structural features (such as concentration, core radius, etc.), thus improving the GC classification performance [25].

In[8] we demonstrated the capability of ML methods to classify GCs using only single band images from Hubble Space Telescope with a classification accuracy of 98.3%, a completeness of 97.8% and only 1.6% of residual contamination. Thus confirming that ML methods may yield low contamination by minimizing the observing requirements and extending the investigation to the outskirts of nearby galaxies.

These results gave us an optimal playground where to train NG models and to validate their potential to solve classification problems characterized by complex data with a noising parameter space.

The paper is structured as follows: in Sect. 2 we describe the data used to test of the various methods. In Sect. 3 we provide a short methodological and technical description of the models. In Sect. 4 we describe the experiments and results about the parameter space analysis and classification experiments, while in Sect. 5 we discuss the results and draw our conclusions.

2 The Astrophysical Playground

As introduced, the HST single band data use dare very suitable to investigate the classification of GCs. They, in fact, are deep and complete in terms of wide-field coverage, i. e. able to sample the GC population, to ensure a high S/N ratio required to measure structural parameters [10]. Furthermore, they provide the possibility to study the overall properties of the GC populations, which usually may differ from those of the central region of a galaxy.

With such data we intend to verify that Neural Gas based models could be able to identify GCs with low contamination even with single band photometric information. Throughout the confirmation of such behavior, we are confident that these models could solve other astrophysical problems as well as in other data-driven problem contexts.

2.1 The data

The data used in the described experiment consist of wide field single band HST observations of the giant elliptical NGC1399 galaxy, located in the core of the Fornax cluster[27]. Due to its distance (D=20.130 Mpc, see[13]), it is considered an optimal case where to cover a large fraction of its GC system with a restricted number of observations. This dataset was used by[25] to study the GC-LMXB connection and the structural properties of the GC population. The optical data were taken with the HST Advanced Camera for Surveys, in the broad V band filter, with 2108 seconds of integration time for each field. The observations were arranged in a 3x3 ACS mosaic with a scale of 0.03 arcsec/pix, and combined into a single image using the MultiDrizzle routine[19]. The field of view of the ACS mosaic covers ~100 square arcmin (Figure 1), extending out to a projected galacto-centric distance of ~55 kpc.

The source catalog was generated using Sextractor [4],[2], by imposing a minimum area of 20 pixels: it contains 12915 sources and reaches 7σ detection at m_V=27.5, i.e. 4 mag below the GC luminosity function, thus allowing to sample the entire GC population (see[8] for details).

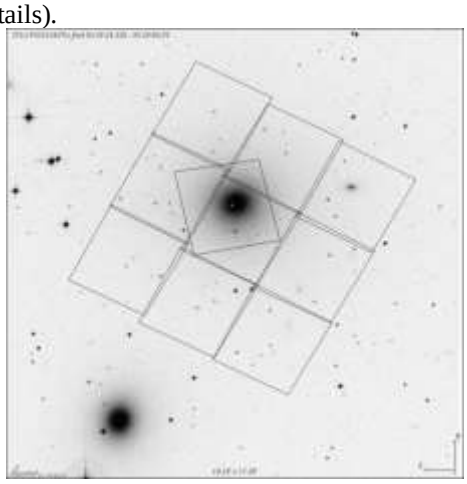


Figure 1 The FoV covered by the HST/ACS mosaic in the broad V band

The source subsample used to build our Knowledge Base (KB) to train the ML models, is composed by 2100 sources with 11features (7 photometric and 4 morphological parameters).

Such parameter space includes three aperture magnitudes within 2, 6 and 20 pixels (*mag_aper1, mag_aper2, mag_aper3*), is ophotal magnitude (*mag_iso*), kron radius (*kron_rad*), central surface brightness (*mu0*), FWHM (*fwhm_im*),and the four structural parameters, respectively, *ellipticity*, King's tidal, effective and core radii (*calr_t, calr_h, calr_c*). The target values of the KB required as ground truth for training and validation, i.e. the binary column indicating the source as GC or not GC, is provided through the typical selection based on multi-band magnitude and colour cuts. The original 2100 sources having a target assigned have been randomly shuffled and split into a training (70%) and a blind test set (30%).

3 The Machine Learning Models

In our work we tested three different variants of the Neural Gas model, using two additional machine learning methods, respectively feed-forward neural network and Random Forest, as comparison benchmarks. In the following all main features of these models are described.

3.1 Growing Neural Gas

Growing Neural Gas (GNG) is presented by[14] as a variant of the Neural Gas algorithm (introduced by[21]), which combines the Competitive Hebbian Learning (CHL, [22]) with a vector quantization technique to achieve a learning that retains the topology of the dataset.

Vector quantization techniques[22] encode a data manifold, e.g. $V \subseteq \mathbf{R}^m$, using a finite set of reference vectors $w = w_1 \dots w_N$, $w_i \in \mathbf{R}^m, i = 1 \dots N$. Every data vector $v \in V$ is described by the best matching reference vector $w_{i(v)}$ for which the distortion error $d(v, w_{i(v)})$ is minimal. This procedure divides the manifold V into a

number of subregions: $V_i = v \in V : \|v - w_i\| \leq \|w - w_j\| \forall j$, called Voronoi polyhedra[24], within which each data vector v is described by the corresponding reference vector w_i.

The Neural Gas network is a vector quantization model characterized by N neural units, each one associated to a reference vector, connected to each other. When an input is extracted, it induces a synaptic excitation detected by all the neurons in the graph and causes its adaptation. As shown in[21], the adaptation rule can be described as a "winner-takes-most" instead of "winner-takes-all" rule:

$$\Delta w_i = \varepsilon h_\lambda(v, w_i)(v - w_i), \qquad i = 1 \dots N \quad (1)$$

The step size ε describes the overall extent of the adaptation. While $h_\lambda(v, w_i) = h_\lambda(k_i(v, w))$ is a function in which k_i is the "neighborhood-ranking" of the reference vectors. Simultaneously, the first and second Best Matching Units (BMUs) develop connections between each other[21].

Each connection has an "age"; when the age of a connection exceeds a pre-specified lifetime T, it is removed[21]. Martinez's reasoning is interesting[22]: they demonstrate how the dynamics of neural units can be compared to a gaseous system. Let's define the density of vector reference at location u through $\rho(u) = F^{-1}_{BMU(u)}$, where $F_{BMU}(u)$ is the volume of Voronoi polyhedra. Hence, $\rho(u)$ is a step function on each Voronoi polyhedra, but we can still imagine that their volumes change slowly from one polyhedra to the next, with $\rho(u)$ continuous. In this way, it is possible to derive an expression for the average change:

$$\langle \Delta w_i \rangle \propto \frac{1}{\rho^{1+\frac{2}{m}}} \left(\partial_u P(u) - \frac{2+m}{m} \frac{P}{\rho} \partial_u \rho(u) \right) \quad (2)$$

where $P(u)$ is the data point distribution.

The equation suggests the name Neural Gas: the average change of the reference vectors corresponds to a motion of particles in a potential $V(u) = -P(u)$. Superimposed on the gradient of this potential there is a force proportional to $-\partial_u \rho(u)$, which points toward the direction of the space where the particle density is low.

Main idea behind the GNG network is to successively add new units to an initially small network, by evaluating local statistical measures collected during previous adaptation steps[14]. Therefore, each neural unit in the graph has associated a local reconstruction error, updated for the BMU at each iteration (i. e. each time an input is extracted): $\Delta error_{BMU} = \|w_{BMU} - v\|$.

Unlike the Neural Gas network, in the GNG the synaptic excitation is limited to the receptive fields related to the Best Matching Unit and its topological neighbors:

$$\Delta w_i = \varepsilon_i(v - w_i), \qquad i \in (BMU, n), \forall\, n \in neighbours(BMU).$$

It is no longer necessary to calculate the ranking for all neural units, but it is sufficient to determine the first and the second BMU.

The increment of the number of units is performed periodically: during the adaptation steps the error accumulation allows to identify the regions in the input space where the signal mapping causes major errors. Therefore, to reduce this error, new units are inserted in such regions[14].

An elimination mechanism is also provided: once the connections, whose age is greater than a certain threshold, have been removed, if their connected units remain isolated (i.e. without emanating edges), those units are removed[14].

3.2 GNG with Radial Basis Function

Fritzke describes an incremental Radial Basis Function (RBF) network suitable for classification and regression problems [14].

The network can be figured out as a standard RBF network [9], with a GNG algorithm as embedded clustering method, used to handle the hidden layer.

Each unit of this hybrid model (hereafter GNGRBF) is a single perceptron with an associated reference vector and a standard deviation. For a given input-output pair $(v, y), v \in \mathbb{R}^n, y \in \mathbb{R}^m$, the activation of the *i*-th unit is described by

$$D_i(v) = e^{-\frac{\|v - w_i\|}{\sigma_i^2}}.$$

Each of the single perceptron computes a weighted sum of the activations:

$$O_i = \sum_j w_{ij} D_j(v), i = 1 \dots m$$

The adaptation rule applies to both reference vectors forming the hidden layer and the RBF weights. For the first, the adaptation rule is the same of the updating rule for the GNG network, while for the weights:

$$\Delta w_{ij} = \eta D_j(y_i - O_i), i = 1 \dots m, j \in N \quad (3)$$

Similarly to the GNG network, new units are inserted where the prediction error is high, updating only the Best Matching Unit at each iteration:

$$\Delta error_{BMU} = \sum_{i=1}^{m} y_i - O_i .$$

3.3 Supervised Growing Neural Gas

The Supervised Growing Neural Gas (SGNG) algorithm is a modification of the GNG algorithm that uses class labels of data to guide the partitioning of data into optimal clusters[15],[20]. Each of the initial neurons is labelled with a unique class label. To reduce the class impurity inside the cluster, the original learning rule (1) is reformulated by considering the case where the BMU belongs or not to the same class of the neuron whose reference vector is the closest to the current input. Depending on such situation the SGNG learning rule is expressed alternatively as:

$$\begin{cases} \Delta w_n = -\varepsilon \dfrac{v - w_n}{\|v - w_n\|} \quad or \\ \Delta w_n = +\varepsilon \dfrac{v - w_n}{\|v - w_n\|} + repulsion(sn, n) \end{cases} \quad (4)$$

Where sn is the nearest class neuron and $repulsion(\square)$ is a function specifically introduced to

maintain neurons sufficiently distant one each other. For the neuron which is topologically close to the neuron sn , the rule intends to increase the clustering accuracy[20]. The insertion mechanism has to reduce not only the intra-distances between data in a cluster, but also the impurity of the cluster. Each unit has associated two kinds of error: an aggregated and a class error. A new neuron is inserted close to the neuron having a highest class error accumulated, while the label is the same as the neuron label with the greater aggregated error.

3.4 Multi Layer Perceptron

The Multi Layer Perceptron **(**MLP**)** architecture is one of the most typical feed-forward neural networks[23]. The term feed-forward is used to identify basic behavior of such neural models, in which the impulse is propagated always in the same direction, e.g. from neuron input layer towards output layer, through one or more hidden layers (the network brain), by combining the sum of weights associated to all neurons.

As easy to understand, the neurons are organized in layers, with proper own role. The input signal, simply propagated throughout the neurons of the input layer, is used to stimulate next hidden and output neuron layers. The output of each neuron is obtained by means of an activation function, applied to the weighted sum of its inputs.

The weights adaptation is obtained by the Logistic Regression rule[17], by estimating the gradient of the cost function, the latter being equal to the logarithm of the likelihood function between the target and the prediction of the model. In this work, our implementation of the MLP is based on the public library Theano[1].

3.5 Random Forest

Random Forest (RF) is one of the most widely known machine learning ensemble methods [5], since it uses a random subset of candidate data features to build an ensemble of decision trees. Our implementation makes use of the public library scikit-learn[26]. This method has been chosen mainly because it provides for each input feature as core of importance (rank) measured in terms of its informative contribution percentage to the classification results. From the architectural point of view, a RF is a collection (forest) of tree-structured classifiers $h(x, \theta_k)$, where the θ_k are independent, identically distributed random vectors and each tree casts a unit vote for the most popular class at input. Moreover, a fundamental property of the RF is the intrinsic absence of training over fitting[5].

4 The experiments

The five models previously introduced have been applied to the dataset described in Sec. 2.1 and their performances have been compared to verify the capability of NG models to solve particularly complex classification problems, like the astrophysical identification of GCs from single-band observed data.

4.1 The Classification Statistical Estimators

In order to evaluate the performances of the selected classifiers, we decided to use three among the classical and widely used statistical estimators, respectively, average efficiency, purity, completeness and F1-score, which can be directly derived from the confusion matrix[28], showed in Figure 2. The *average efficiency*(also known as accuracy, hereafter *AE*), is the ratio between the sum of correctly classified objects on both classes (true positives for both classes, hereafter tp) and the total amount of objects in the test set. The *purity* (als known as precision, hereafter *pur*) of a class measures the ratio between the correctly classified objects and the sum of all objects assigned to that class (i.e. tp/ [tp+fp], where fp indicates the false positives).While the *completeness* (also known as recall, hereafter *comp*) of a class is the ratio tp/ [tp+fn], where fn is the number of false negatives of that class. The quantity tp+fn corresponds to the total amount of objects belonging to that class. The *F1-score* is a statistical test that considers both the purity and completeness of the test to compute the score (i. e. 2 [pur*comp]/ [pur+comp]).

By definition, the dual quantity of the purity is the *contamination*, another important measure which indicates the amount of misclassified objects for each class.

Confusion Matrix	Predicted Class GC	Predicted Class notGC
True class GC	tp	fn
True class notGC	fp	tn

Figure 2 The confusion matrix used to estimate the classification statistics. Columns indicate the class objects as predicted by the classifier, while rows are referred to the true objects of the classes. Main diagonal terms contain the number of correctly classified for the two classes, while fp counts the false positives and fn the false negatives of the GC class

In statistical terms, it is well known the classical tradeoff between purity and completeness in any classification problem, particularly accentuated in astrophysical problems[12]. In the specific case of the GC identification, from the astrophysical point of view, we were mostly interested to the purity, i. e. to ensure the highest level of true GCs correctly identified by the classifiers[8]. However, within the comparison experiments described in this work, our main goal was to evaluate the performances of the classifiers mostly related to the best tradeoff between purity and completeness.

4.2 Analysis of the Data Parameter Space

Before to perform the classification experiments, we preliminarily investigated the parameter space, defined by the 11 features defined in Sec. 2.1, identifying each object within the KB dataset of 2100 objects. Main goal of this phase was to measure the importance of any feature, i.e. its relevance in terms of informative

contribution to the solution of the problem. In the ML context, this analysis is usually called *feature selection*[16]. Its main role is to identify the most relevant features of the parameter space, trying to minimize the impact of the well known problem of the *curse of dimensionality*, i.e. the fact that ML models exhibit a decrease of performance accuracy when the number of features is significantly higher than optimal[18]. This problem is mainly addressed to cases with a huge amount of data and dimensions. However, its effects may also impact contexts with a limited amount of data and parameter space dimension.

The Random Forest model resulted particularly suitable for such analysis, since it is intrinsically able to provide a feature importance ranking during the training phase. The feature importance of the parameter space, representing the dataset used in this work, is shown in Figure 3.

From the astrophysical point of view, this ranking is in accordance with the physics of the problem. In fact, as expected, among the five most important features there are the four magnitudes, i. e. the photometric log-scale measures of the observed object's photonic flux through different apertures of the detector. Furthermore, almost all photometric features resulted as the most relevant. Finally, by looking at the Figure 3, there is an interesting gap between the first six and the last five features, whose cumulative contribution is just ~11% of the total. Finally, a very weak joined contribution (~3%) is carried by the two worst features (*kron_rad* and *calr_c*), which can be considered as the most noising/redundant features for the problem domain.

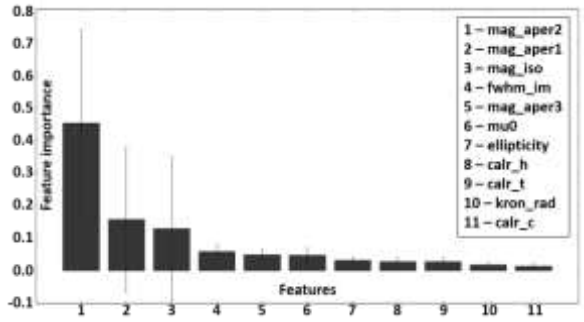


Figure 3 The feature importance ranking obtained by the Random Forest on the 11-feature domain of the input dataset during training (see Sec. 2.1 for details). The blue vertical lines report the importance estimation error bars

Based on such considerations, the analysis of the parameter space provides a list of most interesting classification experiments to be performed with the selected five ML models. This list is reported in Table 1.

The experiment E1 is useful to verify the efficiency by considering the four magnitudes.

The experiment E2 is based on the direct evaluation of the best group of features as derived from the importance results.

The classification efficiency of the full photometric subset of features is evaluated through the experiment E3.

Finally, the experiment E4 is performed to verify the results by removing only the two worst features.

Table 1 List of selected experiments, based on the analysis of the parameter space. The third column reports the identifiers of the included features, according to the importance ranking (see legend in Figure 3)

EXP ID	# features	included features
E1	4	1,2,3,5
E2	6	1,2,3,4,5,6
E3	7	1,2,3,4,5,6,10
E4	9	1,2,3,4,5,6,7,8,9

4.3 The Classification Experiments

Following the results of the parameter space analysis, the original domain of features has been reduced, by varying the number and types of included features. Therefore, the classification experiments have been performed on the dataset, described in Sec. 2.1, composed by 2100 objects and represented by a parameter space with up to a maximum of 9 features (Table 1).

Table 2 Statistical analysis of the classification performances obtained by the five ML models on the blind test set for the four selected experiments. All quantities are expressed in percentage and related to average efficiency (*AE*), purity for each class (*purGC, purNotGC*), completeness for each class (*compGC, compNotGC*) and the F1-score for GC class. The contamination is the dual value of the purity

ID	Estimator	RF %	MLP %	SGNG %	GNGRBF %	GNG %
E1	*AE*	88.9	84.4	88.1	88.1	88.4
	purGC	85.9	80.1	89.7	85.4	83.7
	compGC	87.3	82.6	80.3	85.7	89.2
	F1-scoreGC	86.6	81.3	84.7	85.5	86.4
	purNotGC	91.0	87.6	87.2	90.0	92.1
	compNotGC	89.7	85.6	93.0	89.6	88.1
E2	*AE*	89.0	85.1	87.3	88.3	83.2
	purGC	84.9	77.0	81.0	82.9	74.0
	compGC	89.2	90.7	90.3	90.0	91.1
	F1-scoreGC	87.0	83.3	85.4	86.3	81.7
	purNotGC	92.2	92.6	92.7	92.6	92.6
	compNotGC	89.0	85.6	85.7	87.4	80.0
E3	*AE*	89.0	83.2	85.1	89.2	86.8
	purGC	85.2	77.2	80.0	86.0	84.1
	compGC	88.8	83.8	84.9	88.0	83.8
	F1-scoreGC	87.0	80.4	82.4	87.0	83.9
	purNotGC	91.9	88.0	89.0	91.5	88.7
	compNotGC	89.9	83.2	85.1	89.8	88.4
E4	*AE*	89.5	86.0	88.1	88.7	83.8
	purGC	85.3	82.5	84.1	83.8	78.3

	compGC	90.0	83.8	87.6	90.0	83.8
	F1-scoreGC	87.6	83.1	85.8	86.8	81.0
	purNotGC	92.7	88.6	91.1	92.6	88.1
	compNotGC	89.1	87.5	88.1	88.2	84.1

The dataset has been randomly shuffled and split into a training set of 1470 objects (70% of the whole KB) and a blind test set of 630 objects (the residual 30% of the KB).

These datasets have been used to train and test the selected five ML classifiers. The analysis of results, reported in Table 2, has been performed on the blind test set, in terms of the statistical estimators defined in Sec. 4.2.

5 Discussion and Conclusions

As already underlined, main goal of this work is the validation of NG models as efficient classifiers in noising and multi-dimensional problems, with performances at least comparable to other ML methods, considered "traditional" in terms of their use in such kind of problems.

By looking at Table 2 and focusing on the statistics for the three NG models, it is evident that their result is able to identify GCs from other background objects, reaching a satisfying tradeoff between purity and completeness in all experiments and for both classes. The occurrence of statistical fluctuations is mostly due to the different parameter space used in the four experiments. Nevertheless, none of the three NG models overcome the others in terms of the measured statistics.

If we compare the NG models with the two additional ML methods (Random Forest and MLP neural network), their performances appears almost the same. This implies that NG methods show classification capabilities fully comparable to other ML methods.

Another interesting aspect is the analysis of the degree of coherence among the NG models in terms of commonalities within classified objects. Table 3 reports the percentages of common predictions for the objects correctly classified by considering, respectively both and single classes. On average, the three NG models are in agreement among them for about 80% of the objects correctly classified.

Table 3 Statistics for the three NG models related to the common predictions of the correctly classified objects. Second column is referred to both classes, while the third and fourth columns report, respectively, the statistics for single classes

EXP ID	GC+notGC %	GC %	notGC %
E1	86.0	85.4	86.9
E2	79.8	79.8	79.8
E3	81.1	82.5	79.2
E4	77.8	77.4	78.4

This is also confirmed by looking at the Figure 4, where the tabular results of Table 3 are showed through the Venn diagrams, reporting also more details about their classification commonalities.

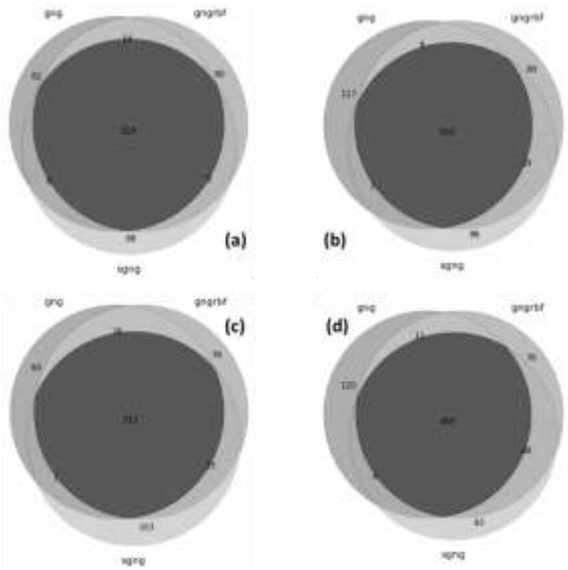


Figure 4 The Venn diagram related to the prediction of all (both GCs and not GCs) correctly classified objects performed by the three Neural Gas based models (GNG, GNGRBF and SGNG) for the experiments, respectively, E1 (a), E2 (b), E3 (c) and E4 (d). The intersection areas (dark grey in the middle) show the objects classified in the same way by different models. Internal numbers indicate the amount of objects correctly classified for each sub-region

Finally, from the computational efficiency point of view, the NG models have theoretically a higher complexity than Random Forest and neural networks. But, since they are based on a dynamic evolution of the internal structure, their complexity strongly depends on the nature of the problem and its parameter space.

Nevertheless, all the presented ML models have a variable architectural attitude to be compliant with the

parallel computing paradigms. Besides the embarrassingly parallel architecture of the Random Forest, the use of optimized libraries, like Theano[1], make also models like MLP highly efficient. From this point of view NG models have a high potentiality to be parallelized. By optimizing GNG, the GNGRBF would automatically benefit, since both share the same search space, except for the RBF training additional cost. In practice, the hidden layer of the supervised network behaves just like a GNG network whose neurons act as inputs for the RBF network. Consequently, with the same number of iterations, the GNGRBF network performs a major number of operations.

On the other hand, the SGNG network is similar to the GNG network, although characterized by a neural insertion mechanism over a long period, thus avoiding too rapid changes in the number of neurons and excessive oscillations of reference vectors. Therefore, on average, the SGNG network computational costs are higher than the models based on the standard Neural Gas mechanism.

In conclusion, although a more intensive test campaign on these models is still ongoing, we can assert that Neural Gas based models are very promising as problem-solving methods, also in presence of complex and multi-dimensional classification and clustering problems, especially if preceded by an accurate analysis and optimization of the parameter space within the problem domain.

Acknowledgements

MB acknowledges the PRIN-INAF 2014 *Glittering kaleidoscopes in the sky: the multifaceted nature and role of Galaxy Clusters*, and the PRIN-MIUR 2015 *Cosmology and Fundamental Physics: illuminating the Dark Universe with Euclid.*

MB, GL and MP acknowledge the H2020-MSCA-ITN-2016 SUNDIAL (*SUrvey Network for Deep Imaging Analysis and Learning*), financed within the Call H2020-EU.1.3.1.

References

[1] Al-Rfou, R., Alain, G., Almahairi, A. et al.: Theano: A {Python} Framework for Fast Computation of Mathematical Expressions. arXiv e-printsabs/1605.02688 (2016)

[2] Annunziatella, M., Mercurio, A., Brescia, M., Cavuoti, S., Longo, G.: Inside Catalogs: A Comparison of Source Extraction Software. PASP 125, 923 (2013). doi: 10.1086/669333

[3] Astroinformatics. In: Brescia, M., Djorgovski, S.G., Feigelson, E.D., Longo, G., Cavuoti, S. (eds.) International Astronomical Union Symposium, 325 (2017). ISBN: 9781107169951

[4] Bertin, E., Arnouts, S.: SExtractor: Software for Source Txtraction. A&A Suppl. Series, 117, pp. 393-404 (1996). doi: 10.1051/aas:1996164

[5] Breiman, L.: Machine Learning, 45. Springer Eds., pp. 25-32 (2001)

[6] Brescia, M., Cavuoti, S., Longo, G., Nocella, A., Garofalo, M., et al.: DAMEWARE: A Web Cyberinfrastructure for Astrophysical Data Mining. PASP. 126, 942 (2014). doi: 10.1086/677725

[7] Brescia, M., Longo, G.: Astroinformatics, Data Mining and the Future of Astronomical Research. Nuclear Instruments and Methods in Physics Research A, 720, pp. 92-94, Elsevier (2013). doi: 10.1016/j.nima.2012.12.027

[8] Brescia, M., Cavuoti, S., Paolillo, M., Longo, G., Puzia, T.: The Detection of Globular Clusters in Galaxies as a Data Mining Problem. MNRAS 421, 2, pp. 1155-1165 (2012). doi: 10.1111/j.1365-2966.2011.20375.x

[9] Broomhead, D.S., Lowe, D.: Radial Basis Functions, Multi-Variable Functional Interpolation and Adaptive Networks. Technical report. RSRE 4148 (1988)

[10] Carlson, M.N., Holtzman, J.A.: Measuring Sizes of Marginally Resolved Young Globular Clusters with the Hubble Space Telescope. PASP 113, 790, pp. 1522-1540 (2001). doi: 10.1086/324417

[11] Cavuoti, S., Garofalo, M., Brescia, M., Paolillo, M., Pescapè, A., Longo, G., Ventre, G.: Astrophysical Data Mining with GPU. A Case Study: Genetic Classification of Globular Clusters. New Astronomy, 26, pp. 12-22 (2014). doi: 10.1016/j.newast.2013.04.004

[12] D'Isanto, A., Cavuoti, S., Brescia, M., Donalek, C., Longo, G., Riccio, G., Djorgovski, S.G.: An Analysis of Feature Relevance in the Classification of Astronomical Transients with Machine Learning Methods. MNRAS 457 (3), pp. 3119-3132 (2016). doi: 10.1093/mnras/stw157

[13] Dunn, L.P., Jerjen, H.: First Results from SAPAC: Toward a Three-dimensional Picture of the Fornax Cluster Core. AJ 132 (3), pp. 1384-1395 (2006). doi: 10.1086/506562

[14] Fritzke, B.: A Growing Neural Gas Network Learns Topologies. In: Advances in Neural Information Processing System, 7, G. Tesauro, D.S. Touretzky and T.K. Leen (eds.), MIT Press, Cambridge MA (1995)

[15] Fritzke, B.: Supervised Learning with Growing Cell Structures. In: Advances in Neural Information Processing System, 6, Cowan, J.D., Tesauro, G., and Alspector, J. (eds.), Morgan-Kaufmann, pp. 255-262 (1994)

[16] Guyon, I., Elisseeff, A.: An Introduction to Variable and Feature Selection. JMLR 3, pp. 1157-1182 (2003)

[17] Harrell, F.E.: Regression Modeling Strategies. Springer-Verlag (2001). ISBN 0-387-95232-2

[18] Hughes, G.F.: On the Mean Accuracy of Statistical Pattern Recognizers. IEEE Transactions on Information Theory, 14 (1), pp. 55-63 (1968). doi:10.1109/TIT.1968.1054102

[19] Koekemoer, A.M., Fruchter, A.S., Hook, R.N., Hack, W.: MultiDrizzle: An Integrated Pyraf Script for Registering, Cleaning and Combining Images. In: The 2002 HST Calibration Workshop. Santiago Arribas, Anton Koekemoer, and Brad Whitmore (eds.). Baltimore, MD: Space Telescope Science Institute (2002)

[20] Jirayusakul, A., Aryuwattanamongkol, S.: A Supervised Growing Neural Gas Algorithm for Cluster Analysis. Springer-Verlag (2006)

[21] Martinez, T., Schulten, K.: A Neural-Gas Network Learns Topologies. In: Artificial Neural Networks. T. Kohonen, K. Makisara, O. Simula, and J. Kangas (eds.), Amsterdam, The Netherlands, Elsevier, pp. 397-402 (1991)

[22] Martinez, T., Berkovich, G., Schulten, K.J.: Neural Gas Network for Vector Quantization and its Application to Time-Series Prediction. In: IEEE Transactions on Neural Networks, 4 (4), pp. 558-569 (1993)

[23] McCulloch, W., Pitts, W., Bulletin of Mathematical Biophysics, 5 (4), pp. 115-133 (1943)

[24] Montoro, J.C.G., Abascal, J.L.F.: The Voronoi Polyhedra as Tools for Structure Determination in Simple Disordered Systems. J. Phys. Chem., 97 (16), pp. 4211-4215 (1993). doi: 10.1021/j100118a044

[25] Paolillo, M., Puzia, T., Goudfrooij, P. et al.: Probing the GC-LMXB Connection in NGC 1399: A Wide-field Study with the Hubble Space Telescope and Chandra. ApJ, 736 (2), p. 90 (2011). doi: 10.1088/0004-637X/736/2/90

[26] Pedregosa, F., Varoquaux, G., Gramfort, A. et al.: Scikit-learn: Machine Learning in Python. JMLR, 12, pp. 2825-2830 (2011)

[27] Puzia, T., Paolillo, M., Goudfrooij, P., Maccarone, T.J., Fabbiano, G., Angelini, L.: Wide-field Hubble Space Telescope Observations of the Globular Cluster System in NGC 1399. ApJ, 786 (2), p. 78 (2014). doi: 10.1088/0004-637X/786/2/78

[28] Stehman, S.V.: Selecting and Interpreting Measures of Thematic Classification Accuracy. Remote Sensing of Environment, 62 (1), pp. 77-89 (1997). doi:10.1016/S0034-4257(97)00083-7

Выявление аномалий в работе механизмов методами машинного обучения

© А.Г. Дьяконов[1] © А.М. Головина[2]

[1] Московский государственный университет имени М. В. Ломоносова,
[2] Московский государственный технический университет имени Н. Э. Баумана,
Москва, Россия

djakonov@mail.ru nastya_gm@mail.ru

Аннотация. Описано исследование по выявлению поломок механизмов методами машинного обучения. Задача сведена к классической задаче машинного обучения без учителя: детектированию аномалий. Сделан обзор современных подходов к решению этой задачи, все они были апробированы на реальных данных. В результате построен алгоритм, который детектирует поломки сложных механизмов в режиме online. В силу своей специфики он также детектирует любое аномальное поведение: нештатный запуск, работу в опасном режиме, неверную эксплуатацию и т. п.

Ключевые слова: большие данные, анализ данных, выбросы, аномалии, поломки.

Anomaly Detection in Mechanisms Using Machine Learning

© A.G. D'yakonov[1] © A.M. Golovina[2]

[1] Lomonosov Moscow State University,
[2] Bauman Moscow State Technical University,
Moscow, Russia

djakonov@mail.ru nastya_gm@mail.ru

Abstract. The research on breakdown detection in mechanisms using machine learning methods is described. The problem is reduced to anomaly detection. The review of modern approaches to anomaly detection is made, all of them have been approved on real data. As a result, the algorithm for online breakdown detection in complicated mechanisms is constructed. By its nature the algorithm also detects any abnormal behavior: emergency start, work in the dangerous mode, incorrect operation, etc.

Keywords: big data, data mining, outliers, anomaly, breakdown.

Введение

В настоящее время стремительно развиваются приложения методов машинного обучения (machine learning) и анализа данных (data mining), что вызвано, с одной стороны, появлением универсальных и практически полезных моделей алгоритмов, например, бустинга (в его современной реализации [7]) и свёрточных нейронных сетей [17], с другой стороны, определённой тенденцией в бизнесе и индустрии улучшать доходность и качество услуг с помощью современных IT-технологий. Отметим, что такая тенденция появилась в последнее десятилетие прежде всего за счёт миниатюризации и удешевления устройств хранения и обработки данных, датификации процессов компаний ([18], постоянного логирования, быстрого перевода в удобный для обработки формат). Такая тенденция привела к появлению специального термина – «Большие данные» (Big Data), как технологии оперирования с современными огромными массивами информации [9].

Если в интернет-компаниях и банках подобные процессы начались раньше (в силу специфики их деятельности, наличия логов, транзакций и т. п.), то в производстве и тяжелой промышленности применение Big Data только начинается.

Прежде всего, здесь возникают следующие группы задач:

- прогноз потребления энергоресурсов и материалов, необходимых для производства, оптимизация закупки и доставки материалов;
- оптимизация процесса производства (построение моделей: как используемые материалы влияют на качество производимого продукта);
- прогнозирование цен на продукцию, прогнозирование спроса, планирование сортамента и оптимизация доставки продукции клиентам;

Труды XIX Международной конференции «Аналитика и управление данными в областях с интенсивным использованием данных» (DAMDID/ RCDL'2017), Москва, Россия, 10–13 октября 2017 года

- диагностика оборудования, обнаружение и прогнозирование неисправностей.

Ниже рассказано о построении алгоритма обнаружения поломок и определения их типа. По договорённости с заказчиком не конкретизируется тип оборудования, на котором проводилась апробация алгоритма, но описана математическая составляющая использованных подходов. Также сделан обзор современных методов обнаружения аномалий (anomaly detection) и результаты их тестирования на данных реальной задачи.

1 Прикладная задача

Массив исследуемых данных состоит из показаний датчиков, установленных на оборудовании. На каждой установке – около 50 датчиков, замеры производятся каждую секунду (такая частота избыточна, поэтому показания были агрегированы до минутных), данные предоставлены за последние 3 года (т. е. около $5 \cdot 10^9$ показаний для одной установки).

Пример показаний датчиков:

- температура (°C);
- давление (кгс/м2);
- уровень шума (дБ);
- скорость вращения (об/мин);
- уровень вибрации (Гц).

На Рис. 1 и 2 показаны примеры сигналов.

Кроме того, есть текстовые логи, в которых описаны, какие работы проводились с оборудованием, а также факты возникновения внештатных ситуаций. Логи вносятся в систему логирования специалистами, поэтому являются текстами с использованием специальных терминов и сокращений. Даты и время внесения в систему добавляются автоматически.

Пример текстового лога

```
10.07.2016 10:03 Нач. бурение 200м 2к скважина станд. р38
10.07.2016 10:48 Кон. бурения Разрыв т15
```

Из логов, в частности, извлекается информация об обнаружении поломок:

- тип поломки;
- где обнаружена (id прибора и секция);
- когда обнаружена;
- когда начались ремонтные работы;
- когда закончены ремонтные работы.

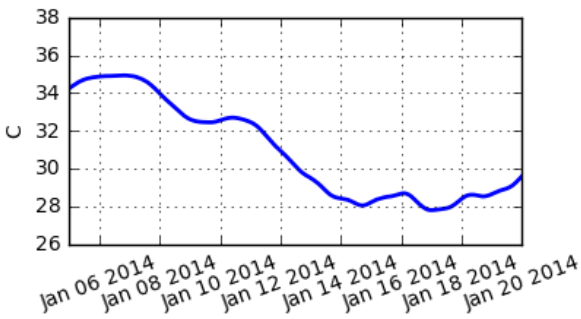


Рисунок 1 Пример сигнала датчика (температура)

За 3 года происходило в среднем 10 поломок на одно оборудование, ремонт длился от 1 дня до 1 месяца.

На основании данных требуется построить алгоритм, который детектирует поломки в режиме online, сообщая тип поломки и её локализацию.

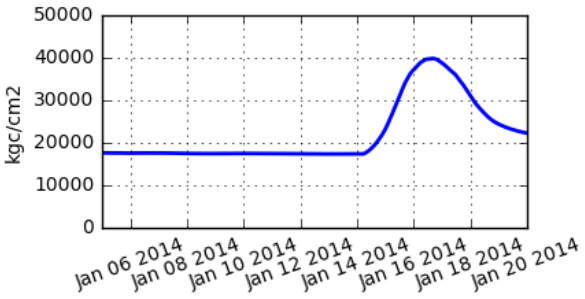


Рисунок 2 Пример сигнала датчика (давление)

2 Обработка текстовых логов

Задача обработки текстовой информации не является центральной при решении описанной выше проблемы детектирования поломок, тем не менее, с помощью обработки получены дополнительные признаки, поэтому опишем её решение.

Для анализа текстовой информации сначала был реализован классификатор текстов, который для каждого лога указывал тип действия, которое ему соответствует: запуск оборудование, выключение, начало ремонтных действий и т. п. Общая схема работы с текстами в рамках данной задачи следующая:

- удаление некоторых спецсимволов;
- переход к буквенным n-граммам;
- модель «мешок слов» (bag of words);
- tf-idf-нормировка;
- решение задач классификации;
- выделение числовых признаков из текстов;

Ниже подробно описаны все этапы.

2.1 Предварительная обработка текстов

Учитывая, что в тексте могут быть сокращения и ошибки (опечатки), изначально текст приводился в один регистр и преобразовывался в буквенные *n*-граммы (большинство спецсимволов удалялось). Лемматизация (приведение к нормальной словарной форме) и удаление стоп-слов не выполнялись [23].

Пример преобразования в 3-граммы

```
Нач. бурение → «НАЧ», «АЧ.», «Ч.Б», «.БУ», «БУР», «УРЕ», «РЕН», «ЕНИ», «НИЕ»
```

Также не были использованы различные методы исправления опечаток. Как показали эксперименты, эту проблему позволяет решить переход к *n*-граммам, поскольку, например, слова «бурение», «буриние» и сокращение «бур.» совпадают по первой триграмме «бур». Кроме того, какие-то отдельные опечатки и сокращения могут соответствовать конкретному человеку, заполнявшему журнал (что позволяет его идентифицировать).

2.2 Перевод текстов в векторную форму

Для перевода текстов в векторную форму

использовался стандартный подход: мешок слов (bag of words, [23]), т. е. была составлена разреженная матрица размера $m \times n$, где m – число текстов, n – число слов во всех текстах, ij-й элемент равен числу, выражающему, сколько раз в i-м тексте встретилось j-е слово. Таким образом, порядок слов в документе не учитывался, а фиксировались лишь числа вхождений слов. Как показали эксперименты, учёт порядка слов переходом к словарным n-граммам не улучшал качество решаемой задачи.

Над построенной матрицей было произведено tf-idf-преобразование [23]. Напомним, что tf-преобразование (term frequency) заключается в вычислении величин

$$\mathrm{tf}(h_{ij}) = \frac{h_{ij}}{\sum_{t=1}^{n} h_{it}}$$

(отношение числа вхождений определённого слова в предложение к числу слов в данном предложении). Смысл подобного преобразования – в инвариантности к повторам текста (скажем, дважды повторенное предложение имеет тот же смысл, что и однократно повторенное).

Idf-преобразование (inverse document frequency) учитывает, что чем чаще встречается слово, тем меньший смысл оно несёт:

$$\mathrm{idf}(h_{ij}) = \frac{m}{\log\left|\{t \mid h_{tj} > 0\}\right|}.$$

Tf-idf-преобразование заключается в замене каждого элемента матрицы на

$$\mathrm{tf}(h_{ij}) \cdot \mathrm{idf}(h_{ij})^{d}.$$

Обычно используется значение степени d=1, но в рассматриваемой задаче почти все слова являются профессиональными терминами, и их частое вхождение не всегда означает бесполезность для решения задачи классификации текста, а степень d как раз контролирует «учёт популярности слов». Было установлено, что оптимальное значение d=0.35.

2.3 Задача классификации текстов

Для «сырых» (необработанных) логов была сделана экспертная разметка: для каждой записи указан тип действия, о котором идёт речь в данной записи. Для разметки использованы логи из первого года, за который есть статистика. Были взяты 1000 случайных записей за год, эксперты выделили 15 классов действий.

Далее был построен классификатор на 15 классов, работа которого в дальнейшем также была оценена экспертами. В качестве тестовой выборки использовались записи 2 и 3 года. Точность определения класса действия – 97% – была признана достаточной.

К сожалению, в этой задаче достаточно трудно сделать экспертную разметку. В отличие от асессорской разметки при поиске, которую может производить практически любой человек, поскольку поисковая выдача как раз и должна быть оптимизирована под нужды среднестатистического пользователя, разметка логов оборудования понятна только людям, знакомым со специальной терминологией.

2.4 Выделение числовых признаков из текстов

Как было показано в примере лога (см. выше), кроме описания действий в нём может содержаться какая-то числовая информация, например, «бурение на глубину 200 метров». Ясно, что число 200 здесь надо уметь вычленять, чтобы потом сравнивать с аналогичными числами в других записях. В задаче классификации числа не учитывались, все они заменялись на специальное слово «number», которое указывало просто наличие какого-то числа в тексте.

Для решения описанной задачи использован следующий подход. Были отобраны типы действий, которые могут содержать числовую информацию. Для каждой числовой информации сформированы правила: где она может встречаться в записи, что ей предшествует и/или что следует после неё. На основе этих правил производился поиск соответствующих чисел. По экспертной оценке точность такого подхода составила 95%.

3 Математическая постановка задачи

После обработки логов изначально заданные многомерные временные ряды показаний датчиков были дополнены категориальными рядами действий, совершаемых с оборудованием (категориальными, поскольку значения ряда в каждый момент времени – тип действия, т. е. категориальная переменная), а также рядами значений признаков, выделенных из текстов (принимают ненулевые значения, только когда совершается соответствующее действие и в логах есть указанное числовое значение, соответствующее этому действию).

Каждый механизм был описан 64-мерным временным рядом. Как отмечено выше, известна информация об обнаружении поломок и их устранении (моменты времени). Необходимо разработать алгоритм автоматического обнаружения поломок: обучить его на статистике первых двух лет и проверить на третьем годе.

Специфика задачи состоит в том, что момент обнаружения поломки не всегда соответствует её возникновению, т. е. поломка может быть обнаружена не вовремя. По оценке экспертов, разность между этими временами может достигать нескольких недель. Кроме того, некоторые поломки могут не оказывать влияния на показания приборов, например, небольшая течь из какого-нибудь шланга (не так сильно изменяются давление и температура, кроме того, изменения происходят плавно).

Описанная задача решалась как задача машинного обучения без учителя – детектирования аномалий (anomaly detection, [5]). Далее представлен обзор современного состояния в области обнаружения аномалий. Главная причина сведения рассматриваемой проблемы именно к этой задаче: основную часть времени оборудование работает в штатном режиме. Поломки выводят оборудование из этого режима: повышается температура отдельных узлов, понижается давление и т. п. Вероятно, статистики достаточно мало, чтобы в значительной степени покрыть все виды поломок, но достаточно,

чтобы описать нормальную работу. Если детектирование аномальной работы будет соответствовать поломкам, то можно использовать такой детектор с требуемой целью.

4 Методы обнаружения аномалий (обзор)

Строго говоря, есть две похожие задачи обнаружения аномалий (Anomaly Detection): детектирование выбросов (Outlier Detection) и «новизны» (Novelty Detection). Как и выброс, «новый объект» – это объект, который отличается по своим свойствам от объектов (обучающей) выборки, но, в отличие от выброса, его в самой выборке пока нет (он появится через некоторое время, задача как раз и заключается в том, чтобы обнаружить его при появлении). Объясним это на примере решаемой задачи. Если по статистике показаний датчиков ищутся моменты времени, когда эти показания сильно отличались от показаний в остальные моменты, то это обнаружение выбросов. Если же статистика используется как пример нормальных показаний, и каждое новое показание проверяется на нормальность (похожесть на старые), то это обнаружение новизны.

Задачи обнаружения аномалий возникают при решении большого числа прикладных проблем, вот далеко не полный их перечень:

- обнаружение подозрительных банковских операций (Credit-card Fraud);
- обнаружение вторжений (Intrusion Detection);
- обнаружение нестандартных игроков на бирже (инсайдеров);
- обнаружение неполадок в механизмах по показаниям датчиков;
- медицинская диагностика (Medical Diagnosis);
- сейсмология.

Далее опишем современные методы обнаружения аномалий.

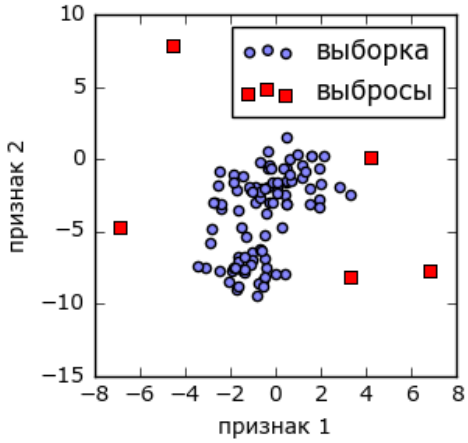


Рисунок 3 Пример выбросов в задаче с двумя признаками

4.1 Статистические тесты

Как правило, их применяют для отдельных признаков и отлавливают экстремальные значения (Extreme-Value Analysis). Для этого используют, например, Z-value [14]:

$$Z_i = \frac{x_i - \mu}{\sigma}$$

или Kurtosis [14]:

$$\frac{1}{n}\sum_{i=1}^{n} Z_i^4.$$

Многие методы визуализации, например, ящик с усами (box plot, [10]), имеют встроенные средства для детектирования и показа таких экстремальных значений.

Важно понимать, что экстремальное значение и аномалия – разные понятия. Например, в небольшой выборке

```
[1, 39, 2, 1, 101, 2, 1, 100, 1, 3, 101, 1, 3, 100, 101, 100, 100]
```

значение 39 можно считать аномалией, хотя оно не является максимальным или минимальным. Также стоит отметить, что аномалия характеризуется, как правило, не только экстремальными значениями отдельных признаков, см. Рис. 3.

4.2 Модельные тесты

Идея очень простая: строим модель, которая описывает данные; точки, которые сильно отклоняются от модели (на которых модель сильно ошибается), и есть аномалии, см. Рис. 4. При выборе модели можно учесть природу задачи, функционал качества и т. п.

Например, в исследуемой задаче можно прогнозировать значения временных рядов с помощью LSTM-нейронной сети [11]. Если реальные значения сильно отличаются от предсказываемых, то это свидетельствует об аномальном поведении. Такой подход хорошо показал себя на недавнем хакатоне лаборатории Касперского [24] по распознаванию аномалий в технологических процессах завода [9].

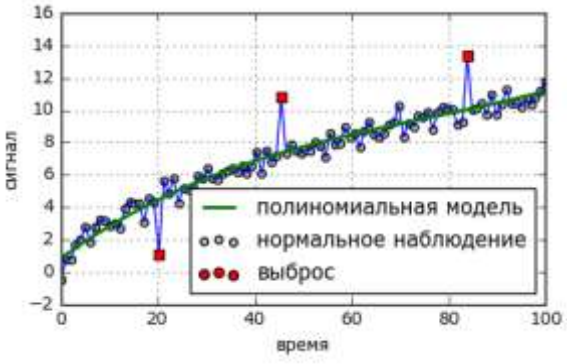


Рисунок 4 Пример применения модельного подхода

Как правило, в задачах обнаружения поломок исходная информация представлена в виде сигналов. Поэтому используется аппарат обработки цифровых сигналов, по крайней мере, на первом этапе решения задачи (для уменьшения размерности и чистки данных). Например, при анализе вибраций [13]

используют дискретное преобразование Фурье (DFT), вейвлеты [21], спектрограммы. В задачах без известной разметки поломок применяют скрытые марковские модели (HMM), а также их различные обобщения [6]. Основная проблема таких алгоритмов – большие временные затраты. Многие подходы практически бесполезны при работе с большими данными из-за использования трудоёмких методов оптимизации: EM-алгоритма [8], MCMC (Markov Chain Monte Carlo) [12] и т. д.

4.3. Итерационные методы

Можно последовательно удалять группы «особо подозрительных объектов». Например, в *n*-мерном признаковом пространстве можно удалять выпуклую оболочку точек-объектов, считая её представителей выбросами, см. Рис. 5. Как правило, методы этой группы достаточно трудоёмки.

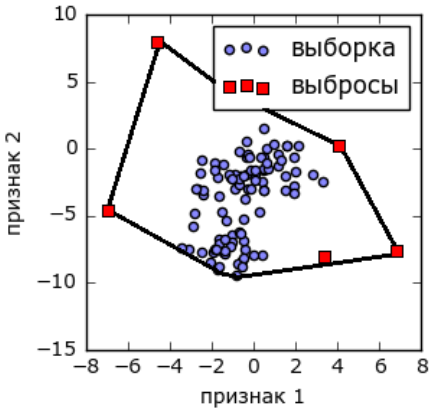


Рисунок 5 Выпуклая оболочка множества точек

4.4. Метрические методы

Это одни из самых популярных методов (судя по числу публикаций, [1]), в них постулируется существование некоторой метрики в пространстве объектов, которая и помогает найти аномалии. Интуитивно понятно, что у выброса мало соседей, а у типичной точки много. Поэтому хорошей мерой аномальности может служить, например, «расстояние до *k*-го соседа» [4]. Здесь используются специфические метрики, например, расстояние Махаланобиса.

4.5. Методы подмены задачи

В этих методах задача обнаружения аномалии заменяется другой задачей, для которой есть удобные и быстрые методы решения. Например, можно сделать кластеризацию, тогда маленькие кластеры, скорее всего, состоят из аномалий, см. Рис. 6.

В исследуемой задаче есть разметка: известны времена обнаружения неисправностей, поэтому описания работы оборудования в эти моменты можно считать классом 1 (размеченные аномалии), а описания работы оборудования после ремонтов и плановых проверок – классом 0 (нормальная работа). Таким образом, решение задачи сводится к решению задачи классификации (classification).

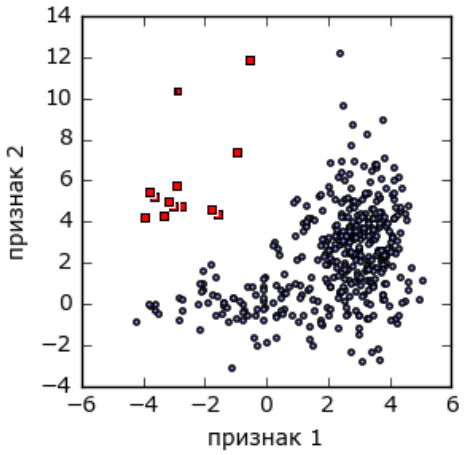


Рисунок 6 Пример конфигурации точек с малым кластером

4.6. Методы машинного обучения

Задачу обнаружения аномалий рассматривают также как отдельную задачу обучения без учителя (unsupervised learning). Такой метод решения может быть отнесён к модельному подходу 4.2, но в этот подраздел вынесены самые популярные алгоритмы (есть реализации в библиотеке scikit-learn языка Python [22]):

- **метод опорных векторов для одного класса** (OneClassSVM, [20]);
- **изолирующий лес** (IsolationForest, [16]);
- **эллипсоидальная аппроксимация данных** (EllipticEnvelope, [19]).

Первый метод – это обычный метод опорных векторов (SVM, [2]), который отделяет выборку от начала координат. Изолирующий лес (Isolation Forest) – это одна из вариаций идеи случайного леса (Random Forest, [3]):

- лес состоит из деревьев;
- каждое дерево строится до исчерпания выборки;
- для построения ветвления в дереве выбираются случайный признак и случайное расщепление,
- для каждого объекта мера его нормальности – среднее арифметическое глубин листьев, в которые он попал (изолировался, см. Рис. 7).

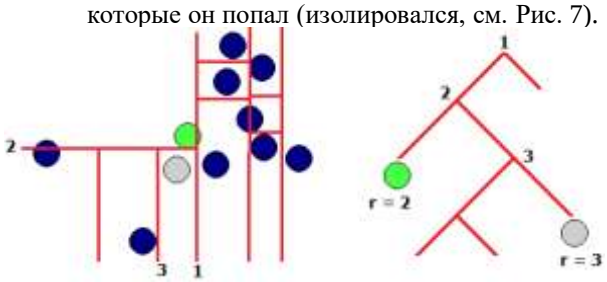


Рисунок 7 Вычисление оценки аномальности в изолирующем лесе

Логика алгоритма простая: при описанном «случайном» способе построения деревьев выбросы

будут попадать в листья на ранних этапах (на небольшой глубине дерева), т. е. выбросы проще «изолировать» (напомним, что дерево строится до тех пор, пока каждый объект не окажется в отдельном листе).

В эллипсоидальной аппроксимации данных, как следует из названия, облако точек моделируется как внутренность эллипсоида. Метод хорошо работает только на одномодальных данных, а особенно хорошо – на нормально распределённых. Степень новизны здесь фактически определяется по расстоянию Махалонобиса.

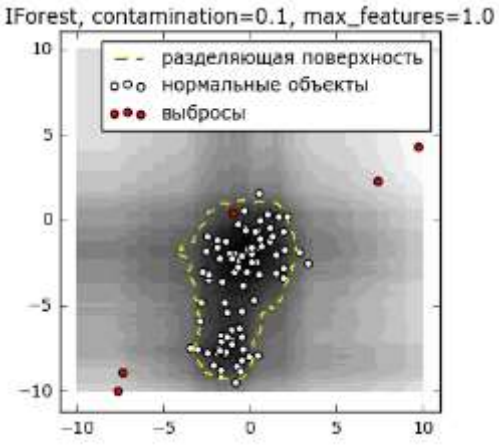


Рисунок 8 Оценка аномальности, полученная с помощью изолирующего леса библиотеки scikit-learn (чем светлее фон, тем аномальнее)

4.7 Ансамбли алгоритмов

Как и во многих других областях машинного обучения, при поиске аномалий часто используют несколько алгоритмов, как правило, разной природы. Каждый из них даёт оценку аномальности, и эти оценки потом «усредняют». Не всегда используют обычное среднее арифметическое, например, иногда хорошее качество показывает максимум (если какой-то алгоритм уверен в аномальности объекта, то, скорее всего, так оно и есть).

Поскольку ключевым моментом в реальных задачах обнаружения аномалий является выбор признаков, которые характеризуют те или иные отклонения от нормы, алгоритмы из ансамбля строят, пытаясь угадать хорошие пространства. Здесь популярны:

- Feature Bagging – для каждого алгоритма берут случайное признаковое подпространство [15];
- Rotated Bagging – в выбранном случайном признаковом подпространстве совершают случайный поворот [1].

5 Исследование методов обнаружения аномалий на реальных данных

Опишем результаты применения различных методов обнаружения аномалий для решения реальной прикладной задачи детектирования поломок. Каждый объект – признаковое описание оборудования в рассматриваемый момент времени. В качестве признаков использовались 64 исходных значения (показания датчиков, а также информация из текстовых логов). Кроме того, были построены признаки, описывающие поведение в прошлом (1 минуту назад, 5 минут, 1 час, 1 сутки назад), и разности между текущими показаниями и показаниями в прошлом. Для некоторых подходов (например, модельного) построения признакового пространства не требуется.

Методы тестировались на последнем году, за который есть статистика. В таблицах указана полнота (какой процент поломок найден) и точность (сколько из детектируемых аномалий действительно являются поломками). Порог детектирования (если оценка анормальности выше него, то алгоритм сигнализирует поломку) подбирался так, чтобы среднее число детектирований совпадало с ожидаемым числом поломок. В таблицах подходы пронумерованы согласно обзору раздела 4.

В целом результаты, представленные в табл. 1, можно считать неудовлетворительными, поскольку заказчик рассчитывал на точность 90% при такой же полноте, но при анализе ошибок были выявлены следующие особенности предложенного подхода. В большинстве случаев детектируется именно аномальное поведение в работе оборудования, т. е. отличающееся от штатного. Поэтому большинство сигналов об анормальности относилось к

- поломкам оборудования;
- неправильной эксплуатации (нарушению правил);
- смене режимов (в том числе, включению и выключению).

В результате получился алгоритм, который детектирует все эти ситуации, что вполне устраивало заказчика. Если пересчитать качество в терминах точности и полноты обнаружения перечисленных ситуаций, то получим Табл. 2.

Таблица 1 Точность и полнота распознавания поломок различными подходами

подход	точность	полнота
4.1	72%	30%
4.2	55%	80%
4.3	68%	32%
4.4	70%	52%
4.5	81%	45%
4.6	80%	80%
4.7	85%	70%

Таблица 2 Точность и полнота распознавания аномалий различными подходами

подход	точность	полнота
4.1	77%	40%
4.2	80%	92%
4.3	70%	45%
4.4	78%	62%
4.5	80%	60%
4.6	97%	87%
4.7	95%	90%

Отметим, что для достижения высокого качества достаточно использовать методы машинного обучения, описанные в разделе 4.6. Использование ансамблей не сильно улучшает качество, но существенно усложняет алгоритмы. Как показано в Табл. 3, самый лучший метод здесь – изолирующий лес.

Таблица 3 Точность и полнота распознавания аномалий методов машинного обучения

метод	точность	полнота
OneClassSVM	88%	85%
IsolationForest	97%	87%
EllipticEnvelope	72%	70%

При правильном детектировании поломки точность определения типа поломки – 87% (для этого решалась отдельная задача классификации), что также оказалось приемлемо, поскольку некоторые типы поломок сложно различать на основе показаний датчиков, например, «низкое давление воды» и «засор подводящих шлангов».

Таблица 4 Среднее время работы алгоритмов разных подходов

подход	время	
	обучения	детектирования
4.1	5 сек	< 1сек
4.2	12 мин	1 сек
4.3	–	9 мин
4.4	9 мин	6 мин
4.5	8 мин	< 1 сек
4.6	10 мин	< 1 сек

В Табл. 4 показано среднее время работы алгоритмов разных групп. Для большинства алгоритмов можно выделить отдельно этап обучения (анализ исходной информации) и детектирования (принятие решения о поломке на основе только что поступивших данных). Все алгоритмы были реализованы на языке Python 3.x.

6 Благодарности

Авторы выражают благодарности компании ООО «Алгомост» за поставленную задачу и консультации со специалистами.

7 Заключение

Разработан алгоритм выявления аномалий в работе оборудования. Кроме поломок, он сигнализирует также о любой некорректной работе и смене режимов работы. Качество оказалось достаточно высоким и полностью удовлетворило заказчика: 97% точности и 87% полноты.

Дальнейшие планы по усовершенствованию алгоритма:

- решение задачи прогнозирования поломок (для заказчика актуально составление расписания проверок и капитального ремонта с учётом износа оборудования);
- решение задачи размещения датчиков (от некоторых датчиков можно отказаться, не снижая качества детектирования поломок);
- использование видеоинформации и изображений (для некоторого оборудования есть кадры съёмки рабочего процесса, которые также регулярно производятся и сохраняются);
- улучшение качества определения типа поломки (пока при решении этой задачи не было сделано такого же масштабного перебора различных подходов, как для детектирования самого факта поломки).

Литература

[1] Aggarwal, C.C.: Outlier Analysis. Springer-Verlag, New York (2013). doi: 10.1007/978-1-4614-6396-2

[2] Boser, B.E., Guyon, I.M., Vapnik, V.N.: A Training Algorithm for Optimal Margin Classifier. Proc. of the Fifth Annual Workshop on Computational Learning Theory – COLT'92, p. 144 (1992). doi: 10.1145/130385.130401

[3] Breiman, L.: Random Forests. Machine Learning, 45 (1), pp. 5-32 (2001). doi:10.1023/A:1010933404324

[4] Breunig, M.M., Kriegel, H.-P., Ng, R.T., Sander, J.: LOF: Identifying Density-based Local Outliers. Proc. of the 2000 ACM SIGMOD Int. Conference on Management of Data, pp. 93-104 (2000)

[5] Chandola, V., Banerjee, A., Kumar, V.: Anomaly detection: A survey. ACM Computing Surveys, 41 (3), pp. 1-58 (2009). doi: 10.1145/ 1541880.1541882

[6] Chao, Y.: Unsupervised Machine Condition Monitoring using Segmental Hidden Markov Models. IJCAI'15 Proc. of the 24th Int Conf. on Artificial Intelligence. AAAI Press, pp. 4009-4016 (2015)

[7] Chen, T., Guestrin, C.: XGBoost: A Scalable Tree Boosting System. Proc. of the 22Nd ACM SIGKDD Int. Conf. on Knowledge Discovery and Data Mining, San Francisco, CA, USA (2016)

[8] Dempster, A.P., Laird, N.M., Rubin, D.B.: Maximum Likelihood from Incomplete Data via the EM Algorithm. J. of the Royal Statistical Society. Series B. 39, pp. 1-38 (1977)

[9] Filonov, P., Lavrentyev, A., Vorontsov, A.: Multivariate Industrial Time Series with Cyber-Attack Simulation: Fault Detection Using an LSTM-based Predictive Data Model. NIPS Time Series Workshop (2016). https://arxiv.org/abs/1612.06676

[10] Frigge, M., Hoaglin, D.C., Iglewicz, B.: Some Implementations of the Box Plot. The American Statistician, 43 (1), pp. 50-54 (1989)

[11] Hochreiter S.: Long Short-term Memory. Neural Computation, 9 (8), pp. 1735-1780 (1997). doi: 10.1162/neco.1997.9.8.1735

[12] Johnson, M.J., Willsky, A.S.: Bayesian Nonparametric Hidden Semi-Markov Models. J. of Machine Learning Research, 14 (1), pp. 673-701 (2013)

[13] Klein, R.: A Method for Anomaly Detection for Non-stationary Vibration Signatures. Annual Conf. of the Prognostics and Health Management Society (2013). https://www.phmsociety.org/sites/phmsociety.org/files/phm_submission/2013/phmc_13_ 038.pdf

[14] Kreyszig, E. Advanced Engineering Mathematics. John Wiley & Sons Inc, 4th edition, 880 p. (1979)

[15] Lazarevic, A., Kumar, V.: Feature Bagging for Outlier Detection. Proc. of the ACM SIGKDD Int. Conf. on Knowledge Discovery and Data Mining, pp. 157-166 (2005). doi: 10.1145/1081870.1081891

[16] Liu, F.T., Tony, T.K.M., Zhou, Z.H.: Isolation Forest. Proc. of the 2008 Eighth IEEE Int. Conf. on Data Mining, pp. 413-422 (2008)

[17] Matusugu, M., Mori, K., Mitari, Y., Kaneda, Y.: Subject Independent Facial Expression Recognition with Robust Face Detection using a Convolutional Neural Network. Neural Networks, 16 (5), pp. 555-559 (2003). doi: 10.1016/S0893-6080(03)00115-1

[18] Mayer-Schönberger, V., Cukier, K.: Big Data: A Revolution that will Transform How We Live, Work, and Think. John Murray, London (2013)

[19] Rousseeuw, P.J., Van Driessen, K.: A Fast Algorithm for the Minimum Covariance Determinant Estimator. Technometrics, 41 (3), pp. 212-223 (1999)

[20] Schölkopf B., et al.: Estimating the Support of a High-dimensional Distribution. Neural Computation, 13 (7), pp. 1443-1471 (2001)

[21] Sheriff, M.Z., Nounou, M.N.: Improved Fault Detection and Process Safety Using Multiscale Shewhart Charts. J. Chem. Eng. Process Technol., 8 (2), pp. 1-16 (2017). doi: 10.4172/2157-7048.1000328

[22] Библиотека алгоритмов машинного обучения для Python, http://scikit-learn.org/stable/

[23] Маннинг, К., Рагхаван, П., Шютце, Х.: Введение в информационный поиск. М.: Изд-во Вильямс (2011)

[24] Хакатон по анализу данных от лаборатории Касперского. https://events.kaspersky.com/hackathon/

Статистическая модель для распознавания смыслов в текстах иностранного языка с обучением на примерах из параллельных текстов

© А.Е. Ермаков © П.Ю. Поляков

ООО «ЭР СИ О»,
Москва, Россия

ermakov@rco.ru pavel@rco.ru

Аннотация. Распознавание смыслов (упоминаний целевых ситуаций, событий и фактов) в текстах иностранного языка в идеале требует разработки синтаксического анализатора этого языка и ряда сопутствующих лингвистических компонентов. В работе предложен альтернативный подход к построению распознавателя смыслов, не требующий глубокого машинного анализа языка текста. Подход строит статистическую модель распознавателя смысла в форме *n*-ок совместно встречающихся слов, с возможностью вставки не более заданного количества посторонних слов между словами *n*-ок. Для обучения модели использованы корпус параллельных текстов и русскоязычный лингвистический анализатор, который выделяет целевые смыслы из русских текстов, отбирая фрагменты, релевантные смыслам, в параллельных текстах иностранного языка. Описаны результаты экспериментов по распознаванию смыслов на корпусе квазипараллельных русско-армянских новостных текстов, в том числе процедура предварительного выравнивания текстов по параллельным фрагментам.

Ключевые слова: машинный анализ текстов на иностранных языках, кросс-языковой информационный поиск, распознавание смысла в тексте, извлечение событий и фактов, статистическое машинное обучение на параллельных текстах, выравнивание параллельных текстов.

Statistical Model for Recognition of Senses in Foreign Language Texts Trained by Examples from Parallel Texts

© Alexander Ermakov © Pavel Polyakov

RCO Llc,
Moscow, Russia

ermakov@rco.ru pavel@rco.ru

Abstract. Recognition of senses (mentioning of target situations, events and facts) in foreign language texts needs developing of a syntactic analyzer and some linguistic components for this language. The alternative approach to construct a senses recognizer that does not need complex machine analysis of the language of a text is proposed in the report. This approach builds a statistical model of a senses recognizer in a form of n-tuples of words that stand together in the text, permitting insertion of a few other words between them. To train the model, a corpus of parallel texts and a Russian linguistic analyzer are applied. The linguistic analyzer is used to extract target senses from Russian texts, selecting the fragments that are relevant to these senses in parallel texts in a foreign language. The results of experiments in senses recognition in the corpus of quasi-parallel Russian-Armenian news texts are described, as well as a preliminary procedure of parallel text fragments alignment.

Keywords: machine analysis of foreign language texts, cross-language information retrieval, recognition of sense in text, events and facts extraction, statistical machine training using parallel texts, parallel texts alignment.

1 Введение

Вопросы межязыкового информационного поиска стали предметом систематического исследования уже с 1990-х годов [2]. Основные результаты и направления современных исследований отражены в работах [3–5, 7]. В центре их внимания оказались статистический машинный перевод, автоматическое построение словарей перевода слов, терминов и именованных сущностей, перевод и расширение поисковых запросов, а также формирование и выравнивание корпусов параллельных текстов-переводов как источников,

Труды XIX Международной конференции «Аналитика и управление данными в областях с интенсивным использованием данных» (DAMDID/ RCDL'2017), Москва, Россия, 10–13 октября 2017 года

необходимых для обучения всех статистических алгоритмов.

В основе предлагаемого нами подхода лежат идеи, имеющие аналогии с таковыми, используемыми в статистическом машинном переводе, наиболее полная информация по которому представлена на веб-ресурсе [7]. Тем не менее, предложенная модель и исследования, посвященные ей, нам не встречались.

Под присутствием заданного смысла в тексте будем понимать описание или упоминание в этом тексте:

- фактов и ситуаций определенного класса, например: *владение акциями предприятий*, *заключение договоров между организациями*, *встречи персон*;
- определенных событий, например: *война в Сирии*, *санкции против России*;
- определенных тем, например: *образ России в зарубежных СМИ, политика Дональда Трампа*.

Тогда задачу информационного поиска в общем виде можно представить как задачу распознавания присутствия заданного смысла в анализируемых текстах и выделения фрагментов текста, релевантных искомому смыслу.

Для распознавания смыслов в русскоязычном тексте можно использовать разработанный нами лингвистический анализатор RCO Fact Extractor [8], который извлекает структурированные описания ситуаций, событий и фактов, выраженные в тексте заданными конфигурациями синтаксически связанных слов [9].

Адаптация русскоязычного лингвистического анализатора к новому языку представляет собой нетривиальную ресурсоемкую задачу, требующую построения синтаксического анализатора этого языка и ряда сопутствующих лингвистических компонентов. В настоящей работе предложен альтернативный подход к построению распознавателя смыслов на иностранном языке, не требующий глубокого машинного анализа этого языка. Подход строит модель статистического распознавателя смысла на новом языке в форме *n*-ок совместно встречающихся слов, с возможностью вставки не более заданного количества посторонних слов между словами *n*-ок. Появление всех слов какой-либо из *n*-ок в пределах текстового окна ограниченной длины интерпретируется как наличие целевого смысла. На практике поиск смысла, описанного в такой форме, может быть эффективно реализован средствами поисковой машины, поддерживающей поиск заданных слов в пределах окна заданной длины с сохранением заданного порядка слов или без такового.

Для обучения распознавателя использованы корпус параллельных текстов и русскоязычный лингвистический анализатор, который выделяет целевые смыслы и содержащие их фрагменты из русских текстов на основе синтактико-семантических шаблонов [9]. Параллельные им фрагменты из текстов иностранного языка также считаются релевантными смыслам и используются для последующей настройки параметров статистической модели. Такой подход требует для настройки распознавателя на каждый новый язык: а) соответствующего параллельного корпуса, представительного в плане присутствия разных способов выражения целевых смыслов; б) простейшего лингвоанализатора, способного строить варианты нормальных форм для словоформ иностранного языка; в) для некоторых видов распознаваемых смыслов от лингвоанализатора может потребоваться умение выделять именованные сущности.

2 Модель статистического распознавателя смыслов

Будем называть *смысло-текстом* текстовый фрагмент, содержащий такую конфигурацию синтаксически связанных слов, появление которой в произвольном тексте говорит о присутствии в нем заданного смысла. Идеальным смысло-текстом является такой фрагмент, в котором отсутствуют лишние слова, появление которых не является обязательным для идентификации присутствия смысла, например: *Берлага заключил договор с Корейко; договор Берлаги и Корейко* (для смысла «*договора между персонами*»); *усиление влияния России на Ближнем Востоке*; *Газпром использует свое монопольное положение на рынке энергоносителей* (смысл «*образ России в зарубежных СМИ*»).

Определим статистический распознаватель смыслов (СРС) как механизм, который для данного текста d определяет, присутствует ли в нем заданный смысл *Se*: формирует реакцию Re(*Se*,*d*)=1, если смысл присутствует, и Re(*Se*,*d*)=0, если отсутствует.

Построим модель СРС в следующем виде. Распознаватель считает, что смысл *Se* присутствует в тексте *d* (реакция Re(*Se*,*d*)=1), если текст содержит хотя бы одну *n*-ку из множества $S=U_{g,n}\ s^g_n$, g=0..G, n=1..N, где $s^g_n=\{(w_1,w_2,\ldots,w_n,g)\}$ – подмножество *n*-ок, каждая из которых содержит n определенных слов w_i, допуская между ними вставку произвольных слов в количестве, не превышающем *g*. Далее будем обозначать профиль СРС как $S=\{s_1,s_2,\ldots,s_J\}$, где J – количество *n*-ок в профиле, нумеруя подряд *n*-ки в профиле и опуская обозначения *n* и *g* в них. Множество *n*-ок *S* будем называть профилем смысла *Se*. В зависимости от степени свободы порядка слов в языке к словам *n*-ок либо следует применять требование сохранения их порядка в окне (пр., армянский, казахский), либо нет (сербский, белорусский). С практической точки зрения достаточными представляются значения N=4, что соответствует, например, упоминанию целевого объекта с тремя дополнительными словами, достаточно точно идентифицирующими искомую ситуацию с объектом.

Обучение СРС смыслу *Se* представляет собой процедуру поиска такого профиля *S*, который обеспечит наилучшее качество работы СРС на

текстах обучающего корпуса D.

За оценку правдоподобия профиля S возьмем совокупную оценку ожидаемых от него полноты P и точности R распознавания смысла (т. н. F_1-мера в теории информационного поиска):

$$q(S,D)=2P(S,D)R(S,D)/(P(S,D)+R(S,D)), \quad (1)$$

где $P(S,D)=|D^*_1(S)|/|D_1(S)|$, $R(S,D)=|D^*_1(S)|/|D(Se)|$, $D(Se)$ – множество смысло-текстов обучающего корпуса, релевантных смыслу Se, $D_1(S)$ – множество всех смысло-текстов, распознанных профилем S, $D^*_1(S)$ – множество смысло-текстов, правильно распознанных профилем S. Тогда наилучший профиль S^*, обеспечивающий максимальное качество СРС, определится как:

$$S^*=\arg\max_S q(S,D) \quad (2)$$

Для ускорения поиска максимума $q(S,D)$ в пространстве S комбинаций n-ок определим правдоподобие вхождения отдельной n-ки s_j в S^* как

$$q(s_j)=(1-1/|D^*_1(s_j)|D^*_1(s_j)|/|D_1(s_j)|, \quad (3)$$

где множитель $|D^*_1(s_j)|/|D_1(s_j)|$ характеризует ожидаемую точность, а множитель $1-1/|D^*_1(s_j)|$ повышает вероятность включения в профиль n-ок с большей частотой встречаемости в релевантных смысло-текстах $D^*_1(s_j)$, поскольку от таковых ожидается большая полнота распознавания смысла. Тогда наилучший профиль S^* в соответствии с (2) можно построить, применив следующий жадный алгоритм поиска в пространстве состояний.

Вначале алгоритм собирает все уникальные n-ки s_j, для которых значение $q(s_j)$ в соответствии с (3) выше определенного порогового значения – кандидатов на включение в профиль. Каждой n-ке-кандидату соответствует массив идентификаторов содержащих ее смысло-текстов d_i.

Далее n-ки сортируются по убыванию значений $q(s_j)$, и первая n-ка включается в профиль на шаге 1: $S_1=\{s_1\}$, чем начинается выполнение итерационного алгоритма расширения профиля новыми n-ми, идя по убыванию значений $q(s_j)$. Обозначим S_{t-1} профиль, полученный на итерации $t-1$, а s_{t-1} – последнюю обработанную n-ку, включенную или не включенную в профиль. На следующей итерации t производится попытка добавить к профилю очередную n-ку s_t. Вычисляются оценки качества нового получаемого профиля $P(S_t,D)$, $R(S_t,D)$, $q(S_t,D)$, и новый профиль S_t признается лучше старого при одновременном соблюдении следующих условий:

$$q(S_t,D)>q(S_{t-1},D) \text{ и } RG(s_t|S_{t-1},D)/TG(s_t|S_{t-1},D)>P_{min}, \quad (4)$$

где P_{min} – минимальная допустимая точность профиля (мы использовали $P_{min}=0.7$), $RG(s_t|S_{t-1},D)$ – прирост количества релевантных смысло-текстов, распознаваемых профилем S_{t-1} после добавления к нему n-ки s_t, $TG(s_t|S_{t-1},D)$ – прирост количества всех смысло-текстов, распознаваемых профилем S_{t-1} после добавления к нему n-ки s_t.

При выполнении обоих условий n-ка добавляется к профилю S_{t-1}, и формируется новый профиль S_t, который принимается за S^*; в противном случае n-ка пропускается, и делается попытка добавления к профилю следующей n-ки s_{t+1} – итерация $t+1$. Расширение профиля прекращается при прохождении всех n-ок-кандидатов или при достижении порога по допустимому количеству n-ок в профиле. Тогда производится возвращение на шаг назад к профилю без добавления последней n-ки, делается попытка добавить следующую за ней n-ку из числа кандидатов и т. д. Таким способом обходится дерево возможных комбинаций n-ок в профиле, и наилучший полученный профиль S^* запоминается. При включении n-ок в порядке убывания их $q(s_j)$ можно ожидать, что лучшие варианты профиля будут получены на более ранних шагах алгоритма.

3 Выравнивание параллельных текстов

Для обучения СРС необходимо сформировать обучающее множество смысло-текстов $D(Se)=\{d_i\}$, релевантных смыслу Se. В качестве таковых отбираются смысло-тексты иностранного языка, параллельные тем русскоязычным смысло-текстам, в которых лингвистическим анализатором выделен смысл Se. Источниками параллельных смысло-текстов, достаточно объемными и представительными в плане разнообразия содержания, являются корпуса переводов новостных сообщений. Такие корпуса в общем случае не содержат строго параллельных текстов, в которых предложения с одинаковыми порядковыми номерами в последовательности могли бы выступать в роли параллельных смысло-текстов. Более того, переводы новостных сообщений часто содержат иную разбивку на предложения, чем их оригиналы, в том числе нередко опускают оригинальные предложения и вставляют новые. Аналогично, перевод предложения может содержать пропуски/вставки ряда значимых слов в описании ситуации – переводчики новостей нередко опускают детали или добавляют собственные интерпретации.

Вследствие этого обучение профилей СРС требует проведения машинной процедуры предварительного выравнивания квазипараллельных текстов, которая устанавливает соответствие между предложениями на двух языках по принципу «одно к одному», «одно к нескольким» или «несколько к одному», а также отбрасывает предложения, перевод которых является излишне «вольным».

Обычно методы выравнивания предложений используют алгоритм динамического программирования, который позволяет вычислительно эффективно определить такую последовательность пар сопоставленных друг другу предложений, для которой сумма расстояний между предложениями в каждой паре будет минимальна. При этом сущность используемого метода заключается в способе определения сходства между парой предложений двух языков. В качестве русскоязычной точки входа в методы выравнивания можно указать работу отечественных исследователей [11]. Наиболее полная информация с зарубежной библиографией по данной теме доступна на веб-ресурсе [6].

Реализованный нами метод требует наличия словаря переводных соответствий слов двух языков, желательно с вариантами синонимичных переводов, а также лингвистических анализаторов обоих языков, способных разделять текст на предложения, а предложения – на сущности, которым приписываются варианты их перевода. В качестве сущностей анализаторы должны выделять слова и, желательно, словосочетания, обозначающие различные классы именованных (персоны, организации, географические объекты) и специальных (даты, периоды времени, денежные суммы) объектов. Именованные и специальные сущности в новостных текстах являются опорными точками для выравнивания параллельных предложений. Сущности приписывается набор альтернативных вариантов перевода (если это удается), а в некоторых случаях, например, для именованных персон, – еще и вариант транскрибирования.

Будем называть количеством сопоставлений переводов $Eq(e_i|d_j)$ количество сущностей из предложения $e_i=\{e^k_i\}$, k=1..K, i=1..I, сопоставленных с сущностями из предложения $d_j=\{d^p_j\}$, p=1..P, j=1..J. Здесь I и J – количества предложений в параллельных текстах E и D; K и P – количества сопоставляемых сущностей в соответствующих предложениях i и j, из числа которых исключены общеупотребимые слова обоих языков, вероятность совпадения переводов которых в паре произвольных предложений высока (прежде всего, это союзы, местоимения, предлоги).

Сущности e^k_i и d^p_j считаются сопоставленными, если выполняется любое из трех условий:

1. обе сущности относятся к классу специальных, и их тип (дата, период времени, денежная суммы) одинаков;
2. один из вариантов имени сущности точно совпадает с одним из вариантов перевода/ транскрипции одного из имен другой сущности;
3. условие 2 выполняется не для точного, а для «нечеткого» совпадения, когда эквивалентными признаются строки, имеющие относительное количество совпавших триграмм символов не менее порогового.

Условие 1 позволяет сопоставить сущности, выражаемые специальными конструкциями (пр. даты), для которых получение совпадающих переводов маловероятно вследствие разнообразия используемых форматов написания в каждом из языков.

Условие 3 необходимо для сопоставления, в первую очередь, именованных сущностей – персон и организаций, при переводе которых человеком-переводчиком часто не соблюдается исходный формат, кроме того, в силу потенциальной неполноты словарей перевода имен, не все части сложных имен могут иметь варианты перевода в словаре. Так, имена персон (как полные, так и краткие) обычно удается сопоставить именно по «нечеткому» совпадению транскрипций. Нередко такое сравнение транскрипций работает для географических мест, обычно не общеизвестных (местных), а также для организаций, напротив, общеизвестных (международных).

Заметим, что величина $Eq(e_i|d_j)$ и вычисляемая наоборот величина $Eq(d_j|e_i)$ в общем случае будут иметь различные значения в силу возможных повторений слов или вариантов их переводов в одном предложении, а также в силу использования «нечеткого» сравнения строк.

Мера прямого сходства переводов определяется как $Tr(e_i|d_j)= Eq(e_i|d_j)$/K, а мера обратного сходства – как $Tr(d_j|e_i)=Eq(d_j|e_i)$/P.

Обозначим $(i(t), j(t))$, t=1...T, последовательность номеров пар предложений e_i и d_j из параллельных текстов $E=\{e_i\}$, i=1..I, и $D=\{d_j\}$, j=1..J, где $j(1)\geq 1$, $i(1)\geq 1$, $j(T)\leq J$, $i(T)\geq I$. Здесь t – переменная, введенная для установления возможного соответствия между номерами предложений $i(t)$ и $j(t)$. Тогда $(i(t),j(t))$ представляет собой возможную последовательность выравнивания предложений при условии, что $i(t)\leq i(t+1)$ и $j(t)\leq j(t+1)$.

В ходе поиска наилучшей последовательности выравнивания $(i(t),j(t))^*$ методом динамического программирования используются два правила:

- пара предложений $(e_{i(t)},d_{j(t)})$ может быть включена в последовательность выравнивания при одновременном выполнении двух условий:
$$\max\{Tr(e_{i(t)},d_{j(t)}),Tr(d_{j(t)}|e_{i(t)})\}>Tr_{\max}$$
и
$$\min\{Tr(d_{j(t)}|e_{i(t)}),\ Tr(e_{i(t)},d_{j(t)})\}>Tr_{\min},$$
где $Tr_{\max}$ и $Tr_{\min}$ – эмпирически подбираемые параметры, в нашем случае – 0.5 и 0.25 соответственно. Увеличение значений $Tr_{\max}$ и $Tr_{\min}$ приводит к повышению точности выравнивания, а их уменьшение – к повышению полноты за счет снижения точности. Чем больше полнота используемого словаря переводных соответствий, тем более высокими могут быть выбраны значения $Tr_{\max}$ и $Tr_{\min}$;
- последовательность выравнивания A признается лучше другой последовательности B, если величина $\sum_t(Eq(e_{i(t)}|,d_{j(t)})+Eq(d_{j(t)}|\ e_{i(t)}))$ – совокупное количество сопоставлений переводов – для последовательности A превышает таковую величину для последовательности B.

После нахождения наилучшего отображения параллельных предложений «одно к одному» делается попытка отобразить предложения, пропущенные в последовательности выравнивания, на те предложения, с которыми уже выровнены предложения, соседние с пропущенными, при реализации выравнивания «одно к нескольким» для случаев несинхронной разбивки исходного и целевого текста на предложения. В контексте задачи обучения CPC процедура выравнивания имеет целью получить смысло-текст минимального размера, поэтому разрешается объединять в один смысло-текст не более двух предложений. В финале происходит отбрасывание тех пар смысло-текстов, для которых мера прямого или обратного сходства переводов оказывается ниже определенного порога –

ожидается, что соответствующий перевод является излишне «вольным».

4 Реализация и эксперименты

Эксперименты по обучению СРС были проведены на корпусе новостных текстов, полученных с армянского сайта http://news.am. Из двух разделов данного сайта (http://news.am/rus/news/ и http://news. am/arm/news/) были скачаны по 300 тысяч русских и армянских текстов, из числа которых по формальному признаку – совпадению идентификаторов – было получено 230 тысяч пар предположительно параллельных русско-армянских текстов.

Для анализа русских текстов был использован лингвистический анализатор RCO Fact Extractor [8], который проводил полный синтаксический анализ текста, выделяя сущности разных типов с отношениями между ними, а также события и факты с их участниками в соответствии с заданными синтактико-семантическими шаблонами [9]. Для анализа армянских текстов был разработан неполный лингвистический анализатор, который разбивал текст на слова и предложения, проводил морфологический анализ и определял для каждого слова возможные варианты его нормальной формы, а также распознавал на основе формальных правил и сворачивал в одну сущность особые цепочки слов – обозначения именованных персон, организаций, географических объектов, дат и обстоятельств времени. Основой для построения армянского морфословаря послужил Восточно-армянский национальный корпус [1], правила описания особых сущностей были разработаны лингвистом на языке Cape для компонента RCO Pattern Extractor [10].

Армяно-русский словарь переводов содержал более 100 тысяч единиц и был сформирован путем консолидации переводов из нескольких интернет-источников. Статистические переводчики Яндекс и Гугл могут переводить по-разному различные словоформы одного и того же слова, например, разным формам армянского слова «ծառայություն» соответствуют формы русских слов *сервис, служба*, *услуга, обслуживание*, а также ряд ошибочных переводов. Эмпирически было подобрано правило определения достоверности переводов, согласно которому признаются недостоверными те варианты, которые встречаются со взвешенной частотой, отношение которой к взвешенной частоте самого частого варианта составляет менее 0.7. Взвешенная частота есть сумма частот встречаемости в каждом из источников, умноженных на вес источника, который определяет уровень доверия к нему. На практике были использованы три источника переводов: а) переводы встретившихся в текстах словоформ, полученные из Яндекса, с весом 1; б) переводы тех же словоформ, полученные из Гугла, с весом 2 (переводы Гугла мы считали достовернее переводов Яндекса); в) строгий словарь объемом 22 тысячи слов (нормальных форм), полученный из интернет-источника http://www.classes.ru/all-armenian/dictionary-armenian-russian.htm, с весом 100, что означало отброс всех вариантов Яндекс- и Гугл-переводов слов, встретившихся в строгом словаре. На переводы в Яндекс и Google были отправлены все армянские словоформы, встретившихся не менее чем в двух документах 230-тысячного корпуса текстов, а также именованные сущности, что составило 350 тысяч единиц.

С использованием полученного словаря переводов алгоритм, описанный в Разделе **3**, разбил 230 тысяч пар текстов на 1370 тысяч пар параллельных фрагментов – смысло-текстов, а для 690 тысяч русских и 585 тысяч армянских предложений не было найдено достаточно близких параллельных переводов. Данная процедура заняла около восьми часов работы одного процессорного ядра.

Программные компоненты обучения СРС работают в три фазы.

На Фазе I обрабатывается корпус xml-файлов, которые формируются двумя лингвистическими анализаторами и содержат описание сущностей, выделенных в армянских смысло-текстах, а также идентификаторы смыслов, которым релевантны параллельные им русские смысло-тексты. Собираются все *n*-ки из нормальных форм сущностей, упоминавшиеся в армянских смысло-текстах, длиной от 2 до 4, допуская встречаемость между словами *n*-ок посторонних слов количеством от 0 до 5. Также собираются параметризованные варианты *n*-ок, в которых конкретные именованные сущности заменяются на свои типы – персона, организация, география. Все омонимичные варианты нормальных форм сущностей порождают соответствующие варианты *n*-ок. Количество разных *n*-ок, получаемых таким образом, имеет порядок сотен миллионов, поэтому для хранения статистики (общие частоты встречаемости *n*-ок в корпусе и частоты *n*-ок по каждому смыслу) в оперативной памяти применяется процедура периодического забывания – как только количество сохраненных *n*-ок превышает 10 миллионов (что не превышает 2 Гбайт ОЗУ), из памяти удаляются данные по наиболее редко встретившимся *n*-кам, имеющим низкие оценки правдоподобия вхождения в профиль какого-либо смысла в соответствии с (3). В финале для каждого смысла отбирается до 1,5 тысяч лучших *n*-ок – кандидатов на последующее включение в профиль, получивших наибольшие оценки правдоподобия вхождения в профиль $q(s_j)$ в соответствии с (3), но не менее 0.01, и сохраняются в файле – препрофиле смысла. Время обработки 230 тысяч новостных текстов для 40 смыслов (см. Таблицу 1) на этой фазе занимает около 4 часов работы одного процессорного ядра.

На Фазе II загружаются файлы препрофилей смыслов, и вновь обрабатывается корпус xml-файлов с описаниями сущностей, выделенных в параллельных смысло-текстах. В результате для каждой *n*-ки в препрофилях подсчитываются частоты ее встречаемости в окнах различной длины с количеством допустимых вставок посторонних слов

от 0 до 5. Одновременно для *n*-ки собираются идентифкаторы смысло-текстов, ее содержащих, по каждому из окон. Собранная информация сохраняется в полных файлах препрофилей смыслов. Время выполнения этой фазы составляет около 1 часа.

На Фазе III загружается файл с полной информацией об *n*-ках препрофилей смыслов и выполняется алгоритм построения профиля СРС, который выбирает *n*-ки из препрофиля в профиль, вычисляя для каждой возможной комбинации n-ок оценку правдоподобия и запоминая комбинацию с максимальной оценкой как лучший вариант профиля S^* в соответствии с (2). Максимальное количество просматриваемых комбинаций ограничивалось 1 миллионом, что оказалось с избытком достаточно для получения наилучшего варианта профиля – средний номер шага процедуры перебора комбинаций, на котором был получен наилучший вариант S^*, по 40 профилям составил около 2 тысяч, а наибольшее из значений (для профиля «путешествия») не превышает 20 тысяч. Для большинства смыслов количество всех комбинаций, подлежащих проверке на выполнение условий (4), оказалось значительно меньше миллиона вследствие относительно небольшого количества обучающих примеров и соответствующих *n*-ок-кандидатов на включение в профиль. В итоге время выполнения данной фазы составило в среднем одну секунду на профиль.

Настройка СРС проводилась на полученном корпусе из 1.370 тысяч пар параллельных смысло-текстов для 40 смыслов – ситуаций, отобранных из более чем 200 типовых ситуаций, распознаваемых русскоязычными лингвистическими шаблонами RCO Fact Extractor. Названия этих смыслов-ситуаций приведены в первом столбце Таблицы 1. Именно к ним обнаружено в корпусе наибольшее количество релевантных смысло-текстов, которое указано в третьем столбце *Exm*.

В экспериментах было построено два отдельных СРС – профили русского СРС строились на русских смысло-текстах и состояли из *n*-ок русских сущностей, выделенных RCO Fact Exctractor, а профили армянского СРС строились на параллельных армянских смысло-текстах и состояли из n-ок армянских слов, выделенных разработанным армянским лингвоанализатором. Обучение СРС «с русского на русский» позволяло исследовать работу СРС в чистом виде, без влияния факторов посторонних составляющих– несовершенств армянского лингвоанализатора, процедуры выравнивания параллельных фрагментов и недостатков собственно параллельных переводов. Значения, полученные для армянского и русского СРС, в Таблице 1 приведены вместе и разделены символом '/'.

Каждая из 40 ситуаций предполагает вовлечение в нее одного или двух участников, представленных в тексте произвольными именованными сущностями. Поэтому, вместо конкретных слов – имен собственных, *n*-ки профилей включали в себя обозначения типов именованных сущностей (O – организация, P – персона, G – географическое место), которые указаны во втором столбце Таблицы 1. Знак '|' разделяет возможные альтернативы. Например, для смысла *владение акциями* во втором столбце указано *O|P O*, что означает, что в *n*-ку слов, входящую в профиль данного смысла, должны обязательно войти какая-либо именованная персона или организация (*владелец акций*) плюс именованная организация (*эмитент акций*).

Различия между цифрами (количество релевантных смысло-текстов) в столбцах *Exm* и *TrainExm* обусловлено следующим. Оценка правдоподобия вхождения *n*-ки в профиль смысла $q(s_j)$ в соответствии с (3) равна 0 в случае единичной частоты встречаемости *n*-ки, вследствие чего такие *n*-ки не могли быть включены в профиль в силу объективной недостаточности данных для обучения СРС. В результате этого многие смысло-тексты, не содержащие ни одной *n*-ки с частотой более 1 и относительно высоким значением $q(s_j)>0{,}01$, фактически не могли участвовать в обучении. Поэтому при расчете значений *R* в соответствии с (1) в качестве $D(Se)$ бралось множество смысло-текстов, содержащих хотя бы одну из *n*-ок-кандидатов на включение в профиль. Это позволяло оценить качество алгоритма обучения относительно независимо от качества обучающей выборки, а также от качества лингвистического анализа армянского текста, которое априори было хуже качества анализа русского – прежде всего, экспериментальный морфоанализатор для армянского языка не мог приводить разные формы слова к одной форме с такой же полнотой и точностью, как морфоанализатор для русского языка. Именно эти фактором в первую очередь обусловлено то, что среднее по столбцу *TrainExm* для армянского языка – 509 – оказалось вдвое меньше, чем для русского – 1056. Соответственно, среднее количество *n*-грамм, включенных в армянские профили, в столбце *n-s* – 145 – оказалось меньше, чем для русского языка – 173. Кроме того, армянские переводы русских новостных текстов нередко опускают описания деталей событий, в которых содержится целевой смысл в исходных текстах, распознаваемый русским лингвоанализатором.

С учетом сказанного средние значения полноты (0.61 для армянского языка против 0.71 для русского в столбце *R*) и точности (0.94 для армянского языка против 0.91 для русского) представляются нам близкими. Соответствующие значения F_1-меры, балансирующей полноту и точность в соответствии с (1), различаются еще меньше – 0.73 против 0.78. Реально ожидаемая полнота, рассчитанная с учетом всех 3587 примеров в корпусе, для русских текстов составляет около 0,21 (0.71 умножить на 1056/3587), а для армянских текстов – 0,09 (0.61 умножить на 509/3587).

Таблица 1 Данные по профилям смыслов. Имена столбцов: Sense – имя смысла; Param – типы сущностей-параметров в *n*-ках; Exm – количество релевантных смысло-текстов в обучающем корпусе; TrainExm – количество релевантных смысло-текстов, участвовавших в обучении профиля; *n*–*s* – количество *n*-ок, вошедших в профиль; P, R – точность и полнота на обучающем корпусе в соответствии с (1). Символом '/' разделены значения, полученные на армянских и русских смысло-текстах

Sense	**Param**	**Exm**	**TrainExm**	**n-s**	**P**	**R**
митинги/забастовки	G	3351	1205/2119	500/500	0.95/0.87	0.62/0.67
уход с рынка	O	25	13/19	5/6	1.0/1.0	0.92/0.74
поставки	O	130	18/71	8/34	0.92/0.90	0.61/0.76
предоставление услуг	O	70	21/43	4/18	1.0/0.88	0.33/0.65
открытие торг. точек	O	224	26/20	11/9	1.0/0.86	0.58/0.95
новые проекты	O	88	34/49	15/21	0.95/0.94	0.56/0.65
проведение тендера	O	108	37/79	10/39	0.96/0.92	0.59/0.82
отзыв продукции	O	131	50/63	23/30	0.94/0.94	0.68/0.78
открытие филиала	O	167	79/122	33/50	0.96/0.88	0.63/0.82
купля/продажа акций	O	437	153/252	58/83	0.98/0.86	0.64/0.88
выпуск товаров	O	549	192/333	61/102	1.0/0.92	0.39/0.52
создание компании	O	565	213/185	51/59	0.98/0.98	0.49/0.68
экономические показатели	O	3212	346/857	182/178	0.97/0.92	0.85/0.69
объединение	O O	222	14/40	1/12	0.83/0.82	0.36/0.80
партнерство	O O	479	79/143	23/53	0.87/0.95	0.58/0.57
рейтинги	O|P	165	29/89	10/29	0.89/0.84	0.55/0.73
юбилей	O|P	90	30/69	10/26	0.89/0.84	0.53/0.86
банкротство	O|P	114	46/70	14/26	0.89/0.92	0.54/0.79
купля/продажа финансов	O|P	750	100/123	36/66	0.93/0.93	0.57/0.78
выигрыш призов	O|P	583	255/374	107/132	0.99/0.91	0.54/0.71
благотворительность	O|P	604	257/285	112/98	0.98/0.91	0.63/0.76
скандалы	O|P	6895	511/1699	166/284	0.94/0.86	0.66/0.65
суды, расследования	O|P	4657	643/2123	93/192	0.95/0.97	0.73/0.60
конфликты	O|P O|P	9932	647/3510	109/380	0.96/0.90	0.72/0.63
финансовая деятельность	O|P	5939	691/1083	230/159	0.94/0.90	0.67/0.72
успехи–неудачи	O|P	5899	1093/2141	331/355	0.89/0.90	0.63/0.72
планы/намерения	O|P	7948	1374/2055	390/273	0.93/0.91	0.54/0.59
мероприятия	O|P	23698	2867/6575	500/500	0.92/0.92	0.61/0.65
владение акциями	O|P O	365	40/95	14/42	0.91/0.99	0.53/0.71
владение организациями	O|P O	2040	431/679	135/183	0.87/0.84	0.69/0.78
договора	O|P O|P	6252	595/1020	211/313	0.93/0.93	0.58/0.67
отставка с должности	P	941	355/521	166/143	0.95/0.92	0.79/0.84
авторство	P	1740	383/555	150/126	0.96/0.92	0.55/0.73
кандидат на выборах	P	2068	529/1020	217/279	0.95/0.91	0.65/0.68
письма	P	2951	819/1102	229/216	0.88/0.87	0.68/0.80
назначение на должность	P	5458	1192/2449	352/417	0.88/0.90	0.72/0.61
путешествия	P G	13295	2564/4304	500/500	0.87/0.86	0.45/0.66
физическое насилие	P P	292	56/142	11/47	1.0/0.95	0.68/0.73
разговор	P P	10216	889/2019	299/436	0.93/0.92	0.55/0.64
встреча	P P	20838	1494/3730	429/500	0.89/0.84	0.73/0.56
среднее		**3587**	**509/1056**	**145/173**	**0.94/0.91**	**0.61/0.71**

5 Заключение

Предложен и экспериментально исследован подход к распознаванию смыслов (упоминаний целевых ситуаций, событий и фактов) в тексте, который допускает относительно простую реализацию предположительно для любого языка, при наличии возможности автоматического выделения требуемых смыслов на русском языке. Подход требует наличия корпуса квазипараллельных текстов – переводов с русского языка на иностранный или обратно. Также желательно наличие простейшего лингвистического анализатора, способного строить варианты нормальных форм для словоформ иностранного языка, что позволяет существенно повысить полноту распознавания смыслов, не требуя примеров параллельных текстов, в которых описывающие смысл слова стоят во всех возможных формах. В зависимости от видов распознаваемых смыслов от лингвистического анализатора может потребоваться умение выделять именованные сущности.

Описанные эксперименты показали высокую точность распознавания смыслов для большого количества разнообразных смыслов (40) на обучающей выборке большого объема (230 тысяч пар квазипараллельных текстов, более 1370 тысяч пар армянских и русских предложений), что, в силу особенностей выбранного способа описания смысла (*n*-ок слов, совместно встречающихся в окне), позволяет ожидать высокой точности распознавания и на других текстах. Невысокая полнота распознавания говорит о необходимости увеличить размер корпуса параллельных новостных текстов в несколько раз (с 230 тысяч пар до миллиона).

В экспериментах не использовалась контрольная выборка текстов, отличная от обучающей, для проверки полученных оценок ожидаемой точности и полноты в силу отсутствия возможности получения качественной экспертной разметки корпуса не только армянских, но и каких-либо других текстов на предмет релевантности различным смыслам. Тем не менее, просмотр содержимого построенных профилей – русских и армянских *n*-ок слов – показал релевантность подавляющего большинства из них целевым смыслам, что повышает уверенность в эффективности подхода.

Литература

[1] Eastern Armenian National Corpus, http://eanc.net

[2] Grefenstette, G. (ed.): Cross-Language Information Retrieval. Springer, 177 p. (1998)

[3] He, D., Wang, J.: Cross-Language Information Retrieval. Information Retrieval: Searching in the 21st Century, Part 11. Wiley and Sons Ltd, pp. 233-254 (2009)

[4] Nie, J-Y.: Cross-Language Information Retrieval. Synthesis Lectures on Human Language Technologies. Morgan & Claypool Publishers, 3 (1), pp. 1-125 (2010)

[5] Nie, J-Y., Gao, J., Cao, G.: Translingual Mining from Text Data. Mining Text Data, Part X. Springer US, pp. 323-359 (2012)

[6] SMT Research Survey Wiki: A Comprehensive Survey of Statistical Machine Translation Research Publications. Sentence Alignment, http:// www.statmt.org/survey/Topic/SentenceAlignment

[7] Statistical Machine Translation, maintained by Philipp Koehn, http://www.statmt.org

[8] RCO Fact Extractor – инструмент компьютерного анализа текстовой информации компании «ЭР СИ О», http://www.rco.ru/?page_id=3554

[9] Ермаков, А.Е., Плешко, В.В.: Семантическая интерпретация в системах компьютерного анализа текста. Информационные технологии, (6), сс. 2-7 (2009)

[10] Ермаков, А.Е., Плешко, В.В., Митюнин, В.А.: RCO Pattern Extractor: компонент выделения особых объектов в тексте. Информатизация и информационная безопасность правоохранительных органов: Сборник трудов XII Межд. науч. конф., Москва, сс. 312-317 (2003)

[11] Потемкин, С.Б., Кедрова, Г.Е.: Выравнивание неразмеченного корпуса параллельных текстов. Компьютерная лингвистика и интеллектуальные технологии: По материалам ежегодной Межд. конф. «Диалог» (Бекасово, 4–8 июня 2008 г.). Вып. 7 (14). М.: РГГУ, сс. 431-437 (2008)

Стендовые и демо презентации

Poster and Demo Session

Individual Optimization of Nutrition on the Basis of Big Data Analysis in Human-Computer Dialogue

© V.N. Krut'ko © N.S. Potemkina © O.A. Mamikonova © A.M. Markova

Federal Research Center "Computer Science and Control" of Russian Academy of Sciences, Moscow, Russia

krutkovn@mail.ru nspotyomkina@mail.ru

Abstract: The paper presents an Internet technology for optimization and supporting an individual nutrition choice on the basis of big data analysis.

Keywords: nutrition optimization, Internet technologies, active longevity.

Introduction

According to WHO about 2/3 of all deaths worldwide are caused by chronic non-communicable diseases (NCDs), including cancer, cardiovascular diseases, type 2 diabetes and others. Most of these deaths occur in low- and middle-income countries, including the Russian Federation. In these countries a progress in NCDs epidemic is observed, which is mostly due to an inadequate nutrition and a refusal of traditional food. Nutritional disorders are also widespread in developed countries, which is manifested in a large percentage of overweight people and in wide spread of NCDs in elderly people [2]. It is a paradox that hunger is not only a lack of food, but also a common situation where an increased caloric value of food is combined with a deficiency of micronutrients. Though the incidence of NCDs is reduced in high-income countries, but it is just high and it is a serious problem. Adequate nutrition can both prevent the development of NCDs and provide therapeutic revival effect [1] and also ensure active longevity.

Characteristics of the problem

At the present stage of the civilization development the person is faced with the following contradictory and complex problems of the nutrition improvement:

1. A contradiction between the requirement of reducing the caloric content associated with a reduction in motor activity and achieved by a reduction of the food weight, and the requirement of a sufficient amount of essential food (nutrient) components, supposing an increase of the food weight.

2. Combinatorial problems in the selection of individual optimal nutrition from the thousands of foodstuffs and BAAs, considering the specific features of the person, his lifestyle, habits, diseases etc. and the complexity of multiple calculations of its elemental composition.

3. A risk of deceptive easiness of solving a problem of a proper nutrition and a compensation for the nutrient lack with BAAs is a digression from the essential requirements of the nutrition density provided with mainly natural products.

4. A nutrition should oftentimes have a preventive, revitalizing, detoxicant and anti-aging function, which means a significant, sometimes multifold, excess of standards on the content of certain vitamins and minerals.

It is impossible to solve these problems without using a conventional nutrition approach. A large amount of data (texts, tables, formulas, equations) determining an individual nutrition selection, requires the use of modern methods for big data analysis, artificial intellect and mathematical optimization.

Structure of the on-line support system for healthy eating

To develop a method for individual computer assessment of nutrition and providing of on-line recommendations for the nutrition selection a systems analysis and a selection of essential interlinks between the user of the developed technology and the environment were performed. The analysis was made considering the following data: condition of the user's health, genetic data, information about lifestyle, psycho-physiological characteristics of the person, information about the daily nutrition, the living conditions, the medical information about the nutrition and scientific based nutrition requirements, in particular.

On the basis of the analysis the main factors that have an impact on the man nutrition were identified, essential interlinks between these factors were found and a concept and a structural model of information-computer support for healthy nutrition was developed.

Using the proposed structural model and previously developed desktop system "Nutrition for Health and Longevity" [3], the system of on-line support for the choice of nutrition is offered. It includes service and content blocks. The latter include:

1. Algorithms of formation of individual standards, considering personal data about psycho-physical load, lifestyle, environment.

Proceedings of the XIX International Conference "Data Analytics and Management in Data Intensive Domains" (DAMDID/RCDL'2017), Moscow, Russia, October 10–13, 2017

2. Subsystem of evaluation of daily individual nutrition, including:

- express evaluation based on a comparison with the Nutrition pyramid ;
- detailed evaluation based on a comparison with the individual norms including about 10–30 nutrients.

3. Subsystem of monitoring, correction planning and optimization of nutrition according to the individual norms.

4. Databases and knowledge bases.

The theoretical concept of web technology is based on a dynamic two-level presentation of information. On the first level there is an analysis of the compliance of individual nutrition with modern scientific nutritional concepts of a proper nutrition shown in the Nutrition pyramid [4]. In the simplest variant with a minimum amount of initial data the users receive individual evaluations and recommendations regarding their nutritional status with the Nutrition pyramid.

The second, more complicated and detailed version of evaluation of the actual individual nutrition provides performing the analysis and the planning of the nutrition based on 20–25 the most important nutrients and requires entering all products used during a certain period. The minimum period is one day, a sufficient one - a week. Based on the entered data the user will get an evaluation of his/her nutrition according to the individual norms and the recommended individual healthy nutrition. Group assessments and recommendations can also be obtained and may be useful for professional users. A concept of two-level data introduction will serve the psychological adaptation of the users: the users of the first level whom the nutrition control will become habitual for, it will be easier to transfer to the work at the second level, which requires more input data.

The most important and interesting feature of the second level of the user data formation is a possibility of using an interactive nutrition optimization procedure. The modern nutrition science considers the nutrition to be an optimal one if it is low-caloric (not exceeding the energy expenditure of modern sedentary person), but it contains a necessary (defined by the norms) amount of all vital nutrients. In addressing the issue of problems of nutrition planning there appears difficult to solve the contradiction between the requirements of the caloric value and the nutrient density of the nutrition. The reduced caloric intake leads to the reduction of the nutritional density, and the attempts to create a nutrition with a necessary nutrient density lead to increasing of the caloric value. This task has no strict mathematical solution and can only be solved by the arrangement of a dialogue between the user and the computer. An interactive dialogue procedure allows the user who is not acquainted with the principles of mathematical optimization, to set the goals and the limits of optimization properly. Thus is effectively implementing the step-by-step instructions for setting and solving the problem of nutrition optimization. However, if the user of the procedure is not a physician, he should be familiar with the basics of nutrition science at the amateur level.

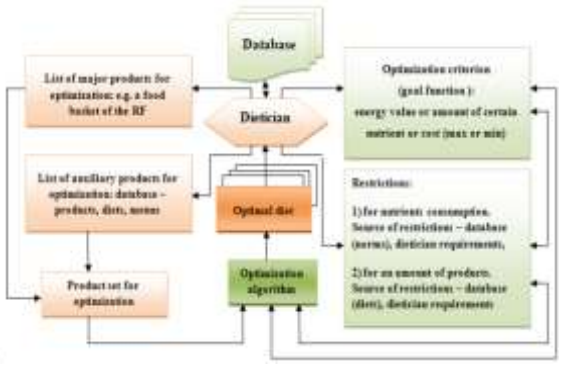


Figure 1 The scheme of interactive optimization

Conclusion

The proposed Web-technology is now in the testing phase. It can be implemented not only on the Internet, but also as a mobile application. It will be useful for both individual and family use, and for the professionals.

Another aspect of great interest is the ability to create and use individual norms, and opportunity to explore and to optimize the professionally developed and officially recommended dietary menus, thus making them balanced and individually adapted.

This work was financially supported by the Ministry of Education and Science of the Russian Federation. Unique Project Identifier RFMEFI60715X0123.

References

[1] Ezzati, M., Riboli, E.: Can Noncommunicable Diseases be Prevented? Lessons from Studies of Populations and Individuals. Science, 337, pp. 1482-1487 (2012)

[2] Mohajeri, MH, Troesch, B, Weber, P.: Inadequate Supply of Vitamins and DHA in the Elderly: Implications for Brain Aging and Alzheimer-type Dementia. Nutrition, 31 (2), pp. 261-275 (2015). doi: 10.1016/j.nut.2014.06.016

[3] Potemkina, N.S.: (2009). Human Ecology, Nutrition Culture and Modern Information Technologies. Aviakosmicheskaia i Ekologicheskaia Meditsina, 43, pp. 13-16. Review (in Russian)

[4] Willett, W.: Eat, Drink, and be Healthy: The Harvard Medical School Guide to Healthy Eating. Simon and Schuster (2011)

Text Categorization Methods Using Topical Importance Characteristic

© Sofia-Nicole Zharikova © Ilya Sochenkov

Peoples' Friendship University of Russia (RUDN University),
Moscow, Russia

sofia-nikol@mail.ru sochenkov_iv@rudn.university

Abstract. This paper presents a study, which evaluates the quality of well-know classification algorithms using Topical Importance Characteristic as a weighting scheme for features. For purposes of research, we used the Twenty Newsgroups dataset. The result of classifiers' performance on different subsets shows that method based on TIC outperforms approaches based on TF-IDF.

Keywords: topical classification, Random Forest classifier, Multinomial Naïve Bayes, Twenty Newsgroups, Topical Importance Characteristic.

1 Introduction

In the modern world there is an issue of information overload. Due to the fact that the amount of news articles and messages is growing, this makes the search of needed information notably complicated. Computer systems were designed to help with this information flood. Thus the basic functions of modern information retrieval software are topical text grouping, near duplicate filtering and text categorization.

The objective of a classification system is to assign documents to predetermined topics. Such classification may be a basis for further analysis, i.e. including topical popularity among users, growing or decreasing amount of messages etc.

In this paper we present a study, which evaluates the quality of some classification algorithms using Topical Importance Characteristic (TIC) as a weighting scheme for features [1]. For purposes of research, we used the Twenty Newsgroups dataset [2][1], and compared the performance of Multinomial naive Bayes (MNB) [3 – 6] and Random Forest classifier (RF) on four schemas for term weighting. We have chosen these machine learning algorithms due to their different nature: MNB is a probabilistic method whilst the RF is the tree-based decision analysis method [7]. So the main research question: is the additional feature information gained with help of TIC worth for classification algorithms?

2 Design of Experiment

In order to control the recall and precision, the Twenty Newsgroups dataset has been divided on two parts: 80% for learning and 20% for testing to perform a five-fold cross-validation.

On the first step of our experiment we extracted specific features from news of the training set and prepare text files for further processing. Later we present each document as a bag of words and compose a dictionary where each word in text will be tied with its unique index which is an integer and then we examine how often this word can be found within the texts. In order to do that we can use the CountVectorizer method from sklearn library [8]. Stop-words (pronouns, prepositions), rare and frequent words were filtered out.

Features were prepared by CountVectorizer, its values of min and max document frequency (DF) were given by formulae (1), (2). It was performed in order to reduce the size of the matrix of features.

$$\min\_df = N \times 0{,}005 \quad (1)$$

$$max\_df = N \times 0{,}22 \quad (2)$$

where *N* is the amount of objects in the training set.

The lengths of texts vary on the dataset, so there are long texts as well as short ones. So the term frequency (TF) seems to be biased. It's obvious that TF of most words would be greater in shorter texts than in longer, even if they are dedicated to the same topic. The TF bursts can bring imbalance to product of linear (TF) and logarithmic (IDF) measures. Thus in this paper we consider some modifications of TF-IDF formula.

Let C to be the universal set of documents (considered as a training dataset); $W(w_i, d)$ – the number of occurrences for the word w_i in text d; $N(d)$ – the total number of occurrences in text d; $A(w_i)$ – the subset of C, which presents documents containing one or more occurrence of the word w_i . Thus, the TF-IDF is defined as follows:

$$tf(w_i, d) = \frac{W(w_i, d)}{N(d)} \quad (3)$$

$$idf(w_i) = \log_2 \frac{|C|}{|A(w_i)|} \quad (4)$$

$$TFIDF(w_i, d) = tf(w_i, d) idf(w_i) \quad (5)$$

Let us introduce the logarithmic term frequency (lTF), normalized IDF, and their product:

$$ltf(w_i, d) = \log_{A(d)+1}(W(w_i, d) + 1) \quad (6)$$

$$idfN(w_i) = \frac{\log_2 \frac{|C|}{|A(w_i)|}}{\log_2 |C|} \quad (7)$$

$$lTFIDF(w_i, d) = ltf(w_i, d)\ idfN(w_i) \quad (8)$$

Let *T* to be a set of topics. For each topic *t* from *T* we have a set of documents related to this topic: *D(t)*. Let $B\big(w_i, D(t)\big)$ to be a number of documents related to the

Proceedings of the XIX International Conference "Data Analytics and Management in Data Intensive Domains" (DAMDID/RCDL'2017), Moscow, Russia, October 10-13, 2017

topic *t* and containing at least one occurrence of the word w_i .

The difference between the information associated with the probability to find at least one occurrence of the word w_i in a text, which is randomly taken from the topic *t*, and a the probability to find at least one occurrence of the word w_i in a text, randomly taken from the rest of C (which is not related to topic *t*) is defined as follows [1]:

$$I(w_i, t, C) = \log_2 \frac{|C \setminus D(t)|}{B(w_i, C \setminus D(t))} - \log_2 \frac{|D(t)|}{B(w_i, D(t))} \quad (9)$$

Using the Heaviside step function H(.) we introduce TIC as follows:

$$TIC(w_i, d, t, C) = ltf(w_i, d)\mathrm{H}\big(I(w_i, t, C)\big) \quad (10)$$

Further we examine two classifiers: RF**Ошибка! Источник ссылки не найден.** and MNB, using the above methods for weighting of features. The usage of such weights for MNB is a little bit tricky and it turns this classifier to a black box of course. But such approach comes from the practice of other researches [4].

Since RF implements the usage of many decision trees, we optimized the total count of trees using the TF-IDF as a weighting scheme. The best accuracy of classification was achieved with 100 trees in RF. For each implemented algorithm we have tested three variants of feature significance evaluation: simple TF-IDF, normalized – lTF-IDF, and based on TIC.

3 Results and discussion

The obtained results are shown in Tables 1-3 using the macro-average of accuracy.

First of all, we have tested classifiers on four categories 'sci.space', 'soc.religion.christian', 'talk.politics.guns', and 'comp.windows.x' of twenty possible. As we can see in the first row, the selected topics significantly differ in meaning, and there are no difficulties for classifier to determine them on a test subset.

An entirely different approach was shown in the second row presenting the accuracy on the topics related to computers: 'comp.os.ms-windows.misc', 'comp.sys.ibm.pc.hardware', 'comp.sys.mac.hardware', 'comp.windows.x'. The accuracy decreases due to the aforementioned topics have a lot of the same words, and all weighting schemas have difficulties with calculating their appropriate weights. Thus the classifiers cannot determine the topics due to the rising noise in features.

Table 1 Accuracy for distant topics

	MNB-TF-IDF	MNB-lTF-IDF	MNB-TIC	RF-TF-IDF	RF-lTF-IDF	RF-TIC
Accuracy (distant topics)	97	97	**99**	96	95	**98**
Accuracy (close topics)	97	97	**99**	85	85	**86**
Accuracy (all topics)	82	80	**95**	80	81	**89**

The third row presents the results for the entire dataset. As we can see, TIC outperforms other competitors on all test cases. RF classifier shows underperformance, while MNB performs well. The analysis of the performance on training and test sets for RF classifier shows that RF suffers a little bit from overfitting.

4 Conclusion

According to the results, we can see that method based on TIC deals with the task better than classic approaches. In single cases normalized lTF-IDF works worse than classic TF-IDF. The overall performance of the proposed approach is quite well.

Support

The research was partially supported by Russian Foundation for Basic Research, project 15-29-06082.

References

[1] Suvorov R., Sochenkov I., Tikhomirov I. Method for pornography filtering in the web based on automatic classification and natural language processing / International Conference on Speech and Computer. – Springer International Publishing, 2013. – P. 233-240

[2] The Twenty Newsgroups. Available: http://qwone.com/~jason/20Newsgroups/ [Accessed 10 June 2017]

[3] Rennie, J.D., Shih, L., Teevan, J., Karger, D.R.: Tackling the poor assumptions of naive Bayes text classifiers. In: Proc Int Conf on Machine Learning. 2003. P. 616–623

[4] Kibriya, A.M., Frank, E., Pfahringer, B., Holmes, G.: Multinomial naive Bayes for text categorization revisited. In: Proc. Australian Conf on AI. 2004 P. 488–499

[5] Frank E., Bouckaert R.R. Naive Bayes for Text Classification with Unbalanced Classes. In: Fürnkranz J., Scheffer T., Spiliopoulou M. (eds) Knowledge Discovery in Databases: PKDD 2006. PKDD 2006. Lecture Notes in Computer Science, vol 4213. Springer, Berlin, Heidelberg. 2006. P. 503-510

[6] Jiang, L., Li, C., Wang, S., & Zhang, L. Deep feature weighting for naive Bayes and its application to text classification. Engineering Applications of Artificial Intelligence, 52, 2016. P. 26-39. doi: 10.1016/j.engappai.2016.02.002

[7] Liaw, A., Wiener, M. Classification and regression by randomForest. R news, 2(3), 2002. P. 18-22

[8] Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., ... & Vanderplas, J. (2011). Scikit-learn: Machine learning in Python. Journal of Machine Learning Research, 12(Oct), 2825-2830

Задачи управления информационно-семантическим полем организации на основе потоковой микросегментации интернет-аудитории

©А.И. Гусева, ©В.С. Киреев, ©П.В. Бочкарев, ©И.А. Кузнецов, ©М.В. Коптелов, ©С.А. Филиппов

Национальный исследовательский ядерный университет «МИФИ»
(Московский инженерно-физический институт)
Россия, Москва

aiguseva@mephi.ru, vskireev@mephi.ru, pvbochkarev@mephi.ru, iakuznetsov@mephi.ru, omoteo@yandex.ru, stanislav@philippov.ru

Аннотация. Управление информационно-семантическим полем организации, включая оценку его текущего состояния на основе анализа тональности сообщений, классификация этих сообщений в соответствии с некоторыми выявленными микросегментами, является крайне актуальной задачей. Данная работа направлена на автоматизацию решения указанной задачи в потоковом режиме с помощью мониторинга пространства социальных сетей и использования методов машинного обучения, таких, как классификационные методы. Приведены результаты экспериментального оценивания информационно-семантического поля АЭС «Куданкулам», показаны результаты анализа тональности, полученные выводы и предложенные изменения в количественные характеристики проекта. Работа выполнена при поддержке Программы повышения конкурентоспособности НИЯУ МИФИ (Договор No 02.a03.21.0005).

Ключевые слова: микросегментация, классификация, оценка тональности, информационно-семантическое поле, наивный байесовский классификатор, атомная отрасль, машинное обучение.

Tasks of the Management of Informational-semantic Field of the Organization on the Basis of the Streaming Micro-segmentation of the Internet Audience

©A.I. Guseva, ©V.S. Kireev, ©P.V. Bochkaryov, ©I.A. Kuznetsov, ©M.V. Koptelov, ©S.A. Philippov

National Research Nuclear University MEPhI (Moscow Engineering Physics Institute)
Russia, Moscow

aiguseva@mephi.ru, vskireev@mephi.ru, pvbochkarev@mephi.ru, iakuznetsov@mephi.ru, omoteo@yandex.ru, stanislav@philippov.ru

Abstract. Management of informational-semantic field of the organization, including the assessment of its current state based on the sentiment analysis of messages, classification of messages in accordance with some identified these micro-segments is a very urgent task. This work is aimed at automating the solution of this problem in the streaming mode using the monitoring space of social networking and the use of machine learning methods such as classification methods. The article presents the results of an experimental evaluation of informational-semantic field of NPP “Kudankulam”, shows the results of the sentiment analysis, the findings and the proposed changes in the quantitative characteristics of the project. This work was supported by the MEPhI Academic Excellence Project (agreement with the Ministry of Education and Science of the Russian Federation of August 27, 2013, project no 02.a03.21.0005).

Keywords: microsegmentation, classification, sentiment analysis, informational-semantic field, naive Bayesian classifier, nuclear industry, machine learning.

1 Введение

Одним из приоритетных направлений Стратегии научно-технологического развития является переход к передовым цифровым, интеллектуальным

Труды XIX Международной конференции «Аналитика и управление данными в областях с интенсивным использованием данных» (DAMDID/ RCDL’2017), Москва, Россия, 10–13 октября 2017 года

технологиям, роботизированным системам, создание систем обработки больших объёмов данных, машинного обучения и искусственного интеллекта. Информационно-семантическое поле организации – это пространство, состоящее из информационных потоков, доступных для анализа и управления, в которых группы потребителей услуг или продуктов, производимых организацией, передают и воспринимают сообщения и сведения о самой организации, продуктах, оказываемых услугах и т. д. Представление об организации для разных групп аудитории отличается в одном и том же информационно-семантическом поле. Выявление причин подобного различия и определение необходимых воздействий на различные группы потребителей (микросегмент) – проблема, решаемая в настоящий момент с помощью больших штатов сотрудников либо нерешаемая вовсе. Целью работы являются исследование и разработка методов микросегментации разнородных групп пользователей в рамках информационно-семантического поля организации и воздействия на микросегменты.

2 Состояние проблемы

В области автоматизированного интеллектуального анализа данных на сегодняшний день существует ограниченное число разработок. Крупнейшим проектом можно считать т. н. проект «Большого механизма» (Big Mechanism), который финансируется агентством DARPA (http://www.darpa.mil/ program/big-mechanism). Кроме этого, можно отметить ряд коммерческих проектов, таких, как Automatic Business Modeler(Algolytics), Automatic Statistician, DataRobot, Quill (Narrative Science), Skytree Machine Learning Software, направленных на создание эффективных предиктивных моделей на основе автоматического анализа сырых данных. Названные проекты специализируются по большей части на количественных данных [2], и основной упор в них сделан на регрессионные модели [1] и первый этап процесса интеллектуального анализа. Также следует выделить исследования, проводимые в отечественном ФИЦ ИУ РАН, направленные на разработку систем обработки больших неструктурированных данных [4], например, многомерно-матричную модель (МОД), позволяющую эффективно использовать параллельные ресурсы в обработке данных.

3 Предлагаемый подход

Предлагаемый подход подразумевает разработку модульной информационной системы (см. Рис. 1).

4 Полученные результаты

Предварительные исследования информационно-семантического поля, проведенные авторским коллективом в области атомной энергетики, а именно, при строительстве объектов атомной энергии за рубежом, показывают значительное влияние данного фактора на результаты проектов [3, 5]. Авторским коллективом впервые был проведен тональный анализ информационно-семантического поля в регионе сооружения атомной станции (АЭС «Куданкулам»). В данном случае был использован подход, основанный на наивном байесовском классификаторе. Было предложено классифицировать все информационные сообщения по определенной шкале и присваивать соответствующее значение индекса информационного риска. Под информационным риском понималась возможность влияния на ход реализации проекта путем изменения его информационно-семантического поля с помощью информационных технологий. Для сбора информации использовалась специализированная система мониторинга СМИ – подсистема интегрированной автоматизированной системы мониторинга угроз ядерно- и радиационно-опасным объектам, которая эксплуатируется в ФГУП «СКЦ Росатома» (предприятие Госкорпорации «Росатом»). Данная система собирает информацию с более 7000 российских и иностранных, новостных и специализированных источников.

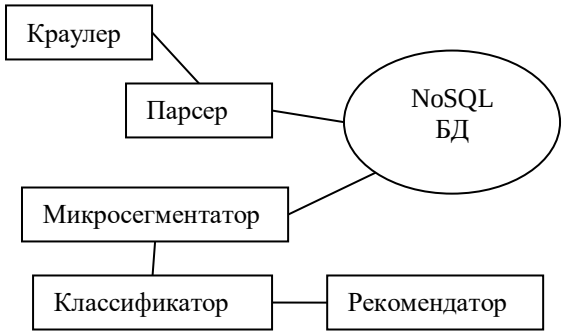


Рисунок 1 Архитектура предлагаемой системы

На основе полученных данных был рассчитан индекс информационного риска Ri (см. Табл. 1), который отражает состояние общественной приемлемости и, как следствие, состояние информационно-семантического поля, для проекта строительства АЭС «Куданкулам» в конкретные периоды времени. Таким образом, сроки сооружения АЭС «Куданкулам» были смещены на 1,5–2 года, что повлекло за собой экономические потери в размере $0,6 млрд. При этом относительный вклад информационного риска, отражающего состояние информационно-семантического поля, в общий риск проекта оказался самым значимым и составил 20%.

Таблица 1 Изменение индекса информационного риска в период сентябрь 2011 г. – август 2013 г.

Временной период	Ri
сентябрь 2011 г. – февраль 2012 г.	0,63
октябрь 2011 г. – ноябрь 2011 г.	0,73
март 2013 г. – август 2013 г.	0,37

5 Заключение

Предлагаемый авторами подход позволяет осуществлять мониторинг и анализ информационно-

семантического поля на систематической основе. Результаты экспериментов, связанные с обследованием АЭС, показывают, что учёт информационного риска может повлечь за собой существенные изменения в сетевом графике работ и стоимости проекта, подтверждая высокую практическую значимость разработанного подхода к микросегментации. Дальнейшими направлениями совершенствования данного подхода являются: использование новых, более точных алгоритмов кластеризации и классификации, а также уточнение шкалы степени информационного риска.

Литература

[1] Lloyd, JR., Duvenaud, D., Grosse, R., Tenenbaum, JB., Ghahraman, Z.: Automatic Construction and Natural-language Description of Nonparametric Regression Models, Association for the Advancement of Artificial Intelligence (AAAI) Conf. (2014)

[2] Kanter, J.M., Veeramachaneni, K.: Deep Feature Synthesis: Towards Automating Data Science Endeavors. Data Science and Advanced Analytics (DSAA). IEEE Int. Conf., pp. 1-10 (2015)

[3] Гусева, А.И., Коптелов, М.В.: Особенности определения риска в инвестиционных проектах строительства АЭС. Атомная энергия, (3 (115)), сс. 170-176 (2013)

[4] Зализняк, А.А.: Лингвоспецифичные единицы русского языка в свете контрастивного корпусного анализа. Компьютерная лингвистика и интеллектуальные технологии: По материалам ежегодной Межд. конф. «Диалог» (Москва, 27–30 мая 2015). Вып. 14 (21). Т. 1: Основная программа конференции. М.: РГГУ, сс. 683-695 (2015)

[5] Коптелов, М.В., Кузнецов, И.А.: Подходы и их реализация при анализе данных общественного мнения о развитии атомного промышленного комплекса. Научное обозрение, (6), сс. 112-114 (2014)

Прозрачный интерфейс для прогноза в машинном обучении

© Ю.О. Кузнецова[1] ©Л.Р. Борисова[2] © А.В. Кузнецова[3,5] © О.В. Сенько[4,5]

[1]Российский национальный исследовательский медицинский университет имени Н.И. Пирогова,
[2]Финансовый университет при Правительстве Российской Федерации,
[3]Институт биохимической физики им. Н.М. Эмануэля РАН,
[4]Федеральный исследовательский центр «Информатика и управление» РАН,
[5]ООО «Азфорус»,
Москва, Россия

jul1998@mail.ru borisovalr@mail.ru azfor@narod.ru senkoov@mail.ru

Аннотация. С использованием метода оптимально достоверных разбиений (ОДР) и модифицированного метода статистически взвешенных синдромов (МСВС) предложен интерфейс прогноза в машинном обучении. Интерфейс позволяет преодолеть проблему «черного ящика»: иллюстрировать процесс прогнозирования с помощью диаграмм рассеяния, ROC-кривой и ранжирования набора информативных показателей с показом расположения исследуемого объекта.

Ключевые слова: машинное обучение, прогнозирование.

Transparent Interface for Prediction in Machine Learning

© Ju.O. Kuznetsova[1] © L.R. Borisova[2] © A.V. Kuznetsova[3,5] © O.V. Senko[4,5]

[1]Pirogov Russian National Research Medical University,
[2]Financial University under the Government of the Russian Federation,
[3]Emanuel Institute of Biochemical Physics of Russian Academy of Sciences,
[4]Federal Research Center Computer Science and Control of the Russian Academy of Sciences,
[5]«Azforus», Ltd,
Moscow, Russia

jul1998@mail.ru borisovalr@mail.ru azfor@narod.ru senkoov@mail.ru

Abstract. A new interface for machine learning predicting models is proposed. Approach is based on optimal valid partitioning (OVP) technique and the modified method of statistically weighted syndromes (LSWR). The interface allows you to overcome the problem of “black box” illustrating prediction process by scatter plots, ROC curves and informative indicators ranking.

Keywords: machine learning, forecasting.

1 Проблема «чёрного ящика» при интерпретации результатов машинного обучения

Активное распространение компьютерных технологий и доступа к интернету в мире привело к удвоению объема информации за последние 2 года. К 2020 г. ожидается увеличение объема до 40 зеттабайт, что превосходит прежние прогнозы на 14%. Растет объем данных, которые потенциально могут быть использованы для решения разнообразных задач диагностики и прогнозирования. Вместе с тем, возрастает роль облачных вычислений в управлении «большими данными» (Big Data). Это также способствует распространению компьютерных методов решения задач диагностики и прогнозирования. Выбор технологии работы с данными зависит от качества и объема данных, поставленной задачи, ограничений по скорости работы и мощности компьютера. Вместе с тем, существенным препятствием для распространения компьютерных технологий является их непрозрачность, т. е. непонятность для конкретного пользователя предлагаемых решений. Такая непрозрачность является существенной в таких областях, как медицина, экономика и др. Ниже

Труды XIX Международной конференции «Аналитика и управление данными в областях с интенсивным использованием данных» (DAMDID/ RCDL’2017), Москва, Россия, 10–13 октября 2017 года

рассмотрены некоторые способы преодоления этой проблемы.

Понятие «черный ящик» печально знакомо специалистам по методам машинного обучения (МО). Большинство таких методов позволяют видеть данные только на входе и на выходе (результат МО), не давая никакого представления, на основании чего данный результат был получен. Мы предлагаем интерфейс для Модифицированного метода «Статистически взвешенных синдромов» [1–4], полностью преодолевший выше обозначенный недостаток. В предлагаемом интерфейсе работа решающего правила при прогнозировании для конкретного объекта полностью прозрачна и наглядна при визуализации на диаграммах рассеяния – на парах показателей, на ROC-кривой и при мониторинге с выводом весов.

Преодоление проблемы «черного ящика» позволяет для прогнозируемого объекта не только получить набор наиболее информативных показателей, позволяющих отнести его к одному из сравниваемых классов, но и видеть прогностические веса в областях многомерного пространства, в которые попадают значения исследуемого объекта. Наглядно можно оценить близость данного значения к основной массе наблюдений объектов обучающей выборки, разделенных на классы с помощью выделения разным цветом. Более того, можно создать план поэтапного перехода из класса с неблагоприятным прогнозом в «хороший» класс, если выделить группу показателей, на которые можно воздействовать с целью их изменения. Ранжирование таких «изменяемых» показателей по информативности и весам позволит в первую очередь обратить внимание на показатели, от которых больше всего зависит результат прогноза для конкретного объекта.

2 Программная система «Прогноз»

В качестве примера рассмотрим дружественный интерфейс для прогностической модели, позволяющей оценить риск повторного случая возникновения коронарного синдрома в первые полгода после перенесенного обострения ишемической болезни сердца (ИБС). Из большого числа известных методов (логистическая регрессия, деревья решений, нейронные сети и др.) отобран подход, давший наибольшую точность прогноза: метод оптимально достоверных разбиений (ОДР) и модифицированный метод статистически взвешенных синдромов (МСВС). AUC=0,72. Подробное описание содержится в [5].

В программу «Прогноз» из 380-ти показателей из базы данных вошел только 41 наиболее информативный показатель. Отбор показателей осуществляли таким образом, чтобы результат на скользящем контроле был наилучшим. Для того чтобы сделать прогноз для конкретного пациента, в таблицу последовательно вводят значения 41 показателя (см. Рис. 1). Для ускорения работы программы предполагается использовать режим быстрого прогноза – без обучения, поскольку ранее полученное при машинном обучении решающее правило работает самостоятельно. Режим с обучением также присутствует и может быть использован, но он требует гораздо большего времени и может быть подключен при пополнении обучающей выборки, в результате чего решающее правило может измениться. В этом случае вывод ROC-кривой также должен быть скорректирован.

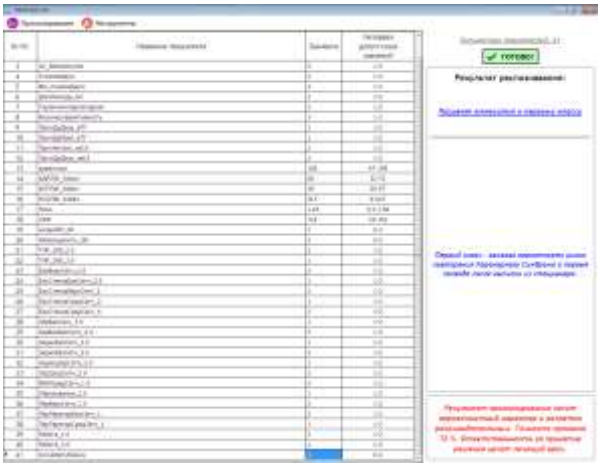


Рисунок 1 Вид интерфейса программы «Прогноз» – ввод данных

Инструменты визуализации представляют собой диаграммы рассеяния (одномерные и двумерные), ROC-кривую и мониторинг, выводящий ранжированные показатели по убыванию информативности. Парный вывод показателей на диаграммы рассеяния с обозначенными границами оптимальных разбиения позволяет видеть достоверное преобладание значений 1 и 2 классов в базовых множествах (квадрантах – областях, ограниченных границами разбиений). Обозначения для классов отличаются цветом.

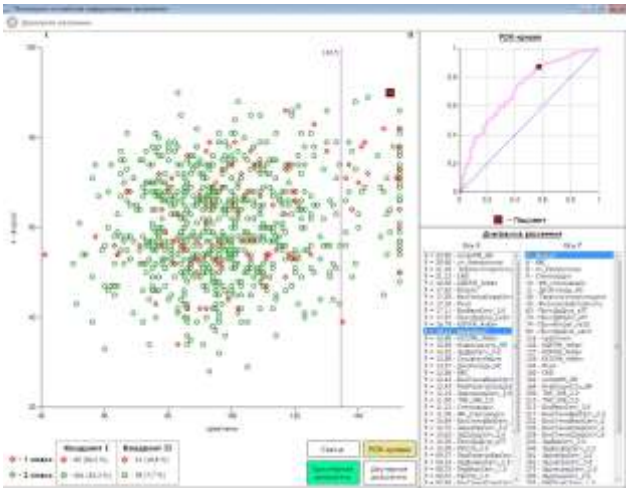


Рисунок 2 Мониторинг, одномерные разбиения. Наиболее информативные показатели

Значения пациента, для которого производится прогноз, выделены на диаграммах более крупным значком, окрашенным в зависимости от класса, к которому пациент отнесен при прогнозировании

(Рис. 2). На ROC-кривой показано аналогичным символом значение для пациента. Чем ближе оно к верхней правой части кривой, тем с большей вероятностью пациент относится к первому классу (неблагоприятный прогноз). Чем ближе положение на кривой к нижнему левому значению, тем прогноз лучше. При изменении значений показателей – в динамике, результат прогноза может также изменяться. В этом случае можно будет видеть перемещение положения символа для пациента вдоль ROC-кривой.

Мониторинг позволяет также закрашивать базовые множества в зависимости от достоверного преобладания там объектов 1-го или 2-го классов. Под диаграммой приведено количество объектов по классам в каждом базовом множестве и их процентное соотношение. В случае попадания значения пациента, для которого производится прогноз, в базовое множество с преобладанием значений показателей объектов 1 класса можно видеть, насколько удален символ пациента от границы, разделяющей два класса.

Для базового множества, в которое попало значение для исследуемого объекта, выводится значение веса для 1 класса. Ранжирование по весам в базовых множествах позволяет также обратить внимание на наиболее важные сочетания показателей, влияющие на результат прогнозирования.

Вывод отчета в файл происходит по нажатию одной кнопки. Список информативных показателей, ранжированных по функционалу (Х-квадрат), и весов за 1 класс по базовым множествам сохраняется в документ. Могут быть выведены и все необходимые диаграммы рассеяния.

Заключение

Аналогичные программные интерфейсы для прогнозирования могут помочь при машинном обучении в других сферах: в экономике, маркетинге, банковском скоринге и прочих. Например, в задачах бенчмаркетинга требуется отслеживание параметров развития предприятия, а также оценки его в ряду аналогичных компаний. Необходимо проводить поиск наиболее эффективных решений при конкуренции с предприятиями, имеющими успех в аналогичных условиях. Для этого аналитикам при машинном обучении на собранной информации важно не только выяснить сходства и различия в работе исследуемого предприятия и «образцов»-аналогий, но и выявить причины отставания, выделить полезный опыт, что и готов предоставить предлагаемый интерфейс: будут выявлены и наглядно представлены группы наиболее близких по своим показателям компаний-конкурентов и будут выделены наиболее значимые показатели, по которым необходимо произвести изменения в первую очередь для перехода в более успешную группу компаний.

Программа интерфейса «Прогноз» полностью подготовлена для работы с другими данными. При постановке новой задачи прогнозирования адаптация алгоритма к новой обучающей выборке займет минимум времени. Удобство использования предлагаемого интерфейса ввиду его наглядности и полной прозрачности в значительной степени улучшит понимание аналитиком закономерностей, выявленных в процессе машинного обучения. Это преимущество в свою очередь позволит понять наиболее значимые процессы, приводящие к тому или иному состоянию системы, а также пути перехода в нужное положение в многомерном пространстве признаков.

Благодарности

Работа была выполнена при поддержке РФФИ, грант 17-07-01362.

Литература

[1] Senko, O.V., Kuznetsova, A.V.: A recognition method based on collective decision making using systems of regularities of various types. Pattern Recognition and Image Analysis, 2 (2), pp. 152-162 (2010)

[2] Kuznetsova, A.V., Kostomarova, I.V., Senko, O.V.: Modification of the Method of Optimal Valid Partitioning for Comparison of Patterns Related to the Occurrence of Ischemic Stroke in two Groups of Patients. Pattern Recognition and Image Analysis, 24 (1), pp. 114-123 (2014)

[3] Senko, O.V., Kuznetsova, A.V. The Optimal Valid Partitioning Procedures. InterStat, (2) (2006)

[4] Кузнецов, В.А., Сенько, О.В., Кузнецова, А.В., Семенова, Л.П., Алещенко, А.В., Гладышева, Т.Б., Ившина, А.В.: Распознавание нечетких систем по методу статистически взвешенных синдромов и его применение для иммуногематологической нормы и хронической патологии. Химическая физика, 15 (1), сс. 81-100 (1996)

[5] Гулиев, Р.Р., Сенько, О.В., Затейщиков, Д.А. и др.: Применение оптимальных разбиений для многопараметрического анализа данных в клинических исследованиях. Математическая биология и биоинформатика, 11 (1), сс. 46-63 (2016)

Указатель авторов

M

N

O

P

Q

R

S

T

U

V

Z

А

Б

Научное издание

АНАЛИТИКА И УПРАВЛЕНИЕ ДАННЫМИ
В ОБЛАСТЯХ С ИНТЕНСИВНЫМ ИСПОЛЬЗОВАНИЕМ ДАННЫХ

Сборник научных трудов
XIX Международной конференции DAMDID / RCDL'2017
10-13 октября 2017 г.
г. Москва, МГУ, Россия

Под редакцией Л.А. Калиниченко, Я. Манолопулос, Н.А. Скворцова, В.А. Сухомлина

DATA ANALYTICS AND MANAGEMENT
IN DATA INTENSIVE DOMAINS

Collection of Scientific Papers of the
XIX International Conference DAMDID / RCDL'2017
October 10-13, 2017
Moscow, MSU, Russia

Edited by L.A. Kalinichenko, Y. Manolopoulos, N.A. Skvortsov, V.A. Sukhomlin

Подписано в печать 25 сентября 2017 года
Формат 60x84 1/8. Бумага офсетная. Печать цифровая.
Усл.-печ. л. 61,5. Уч. -изд. л. 55,2.
Тираж 160 экз.
Заказ № 119960.

Отпечатано в типографии
ПАО «Т8 Издательские Технологии»
109316, г. Москва, Волгоградский
проспект, д. 42, корп. 5, ком. 6
+7 (499) 322-38-30
info@t8print.ru
www.t8print.ru