Федеральный исследовательский центр «Информатика и управление» РАН
Национальный исследовательский университет Высшая школа экономики
Национальный исследовательский ядерный университет «МИФИ»
Московская секция ACM SIGMOD
Российский фонд фундаментальных исследований

Аналитика и управление данными в областях с интенсивным использованием данных

XVIII Международная конференция DAMDID / RCDL'2016

Ершово, Московская обл., 11–14 октября 2016 года

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

Data Analytics and Management in Data Intensive Domains

XVIII International Conference DAMDID / RCDL'2016

October 11–14, 2016, Ershovo, Moscow Region, Russia

ТОРУС ПРЕСС МОСКВА 2016

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

А 64 Аналитика и управление данными в областях с интенсивным использованием данных: XVIII Международная конференция DAMDID / RCDL'2016 (11–14 октября 2016 года, Ершово, Московская обл., Россия): труды конференции – М.: ФИЦ ИУ РАН, 2016. 428 с.

ISBN 978-5-94588-206-5

Конференция «Аналитика и управление данными в областях с интенсивным использованием данных» ("Data Analytics and Management in Data Intensive Domains", DAMDID) 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, медицина, физика, химия, и пр.), социальных наук, а также различных ОИИД информатики, 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», http://rcdl.ru) с сохранением преемственности по отношению к RCDL после многих лет ее успешного функционирования.

ББК [73+32.973.233]я431

Data Analytics and Management in Data Intensive Domains: 18th International Conference DAMDID / RCDL'2016 (October 11–14, 2016, Ershovo, Moscow Region, Russia): Conference Proceedings. – Moscow: FRC CSC RAS, 2016 – 428 p.

ISBN 978-5-94588-206-5

The "Data Analytics and Management in Data Intensive Domains" conference (DAMDID) is planned traditionally 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, medicine, 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 arranged 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 978-5-94588-206-5

XVIII Международная конференция DAMDID / RCDL’2016

Аналитика и управление данными в областях с интенсивным использованием данных

11-14 октября 2016 года, Ершово, Москва

http://damdid2016.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 (ВМК МГУ, Москва)
Елизаров Александр Михайлович (Приволжский федеральный университет, Казань)
Захаров Виктор Николаевич (ИПИ 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)
Литвин Владимир Андреевич (Evogh Inc, CalTech, США)
Майсурадзе Арчил Иериевич (МВК 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’2016

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

Организационный комитет

Сопредседатели

Борис Позин (МИЭМ НИУ ВШЭ)

Виктор Захаров (ФИЦ ИУ РАН)

Александр Невзоров (НИЯУ МИФИ)

Члены оргкомитета

Дмитрий Брюхов (ФИЦ ИУ РАН), поддержка Веб сайта

Дмитрий Ковалев (ФИЦ ИУ РАН), контакты, переписка

Николай Скворцов (ФИЦ ИУ РАН), поддержка CMT, редактирование материалов конференции

Евгения Дударева (ФИЦ ИУ РАН), организационно-финансовые вопросы

Анна Надеждина (ФИЦ ИУ РАН), казначей

Юлия Трусова (ФИЦ ИУ РАН), визовая поддержка

Татьяна Дударева (ПРОТЭЙ ТРЕВЭЛ), председатель локального обустройства

Сергей Верещагин (ИНАСАН), связь оргкомитета с д/о Ершово

Протей тревел, д/о Ершово, транспорт для участников конференции

МИЭМ НИУ ВШЭ, д/о Ершово, обеспечение волонтеров, техническая поддержка, включая компьютеры, проекторы, Wi-Fi в аудиториях и жилых помещениях

МИФИ, синхронный перевод

Программный комитет

Сопредседатели

Леонид Калиниченко (ФИЦ ИУ РАН)

Сергей Кузнецов (НИУ ВШЭ)

Яннис Манолопулос (Университет им. Аристотеля в Салониках)

Члены программного комитета

Карл Аберер (EPFL, Лозанна, Швейцария)

Пламен Ангелов (Университет Ланкастера, Великобритания)

Аркадий Авраменко (Пущинская радиоастрономическая обсерватория)

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

Павел Браславский (Уральский федеральный университет, СКБ Контур, Екатеринбург)

Василий Бунаков (STFC, Харвелл, Оксфордшир, Великобритания)

Наталья Васильева (ЗАО «Хьюлетт-Паккард АО»)

Питер Виттенбург (Институт Макса Планка)

Алексей Вовченко (ФИЦ ИУ РАН, Москва)

Вячеслав Вольфенгаген (НИЯУ «МИФИ», Москва)

Константин Воронцов (ВМК МГУ, Москва)

Домагой Вргоч (Центр исследований Семантического Веба)

Евгений Гордов (ИМКЭС СО РАН)

Ольга Горчинская (ФОРС, Москва)

Максим Губин (Google)

Эрнесто Дамиани (Khalifa University, Абу-Даби)

Юрий Демченко (Университет Амстердама)

Шломи Долев (DCS, Ben-Gurion University of the Negev, Беэр-Шева, Израиль)

Борис Добров (НИВЦ МГУ им. М.В. Ломоносова)

Александр Елизаров (Казанский федеральный университет)

Олег Жижимов (Институт вычислительных технологий СО РАН)

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

Виктор Захаров (ИПИ ФИЦ ИУ РАН)

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

Панос Крисантис (Университет Питтсбурга)

Пол Коэн (Школа информации, Университет Аризоны)

Леонид Калиниченко (ИПИ ФИЦ ИУ РАН)

Джордж Карипис (Университет Миннесоты, Миннеаполис)

Надежда Киселева (ИМЕТ РАН)

Алексей Климентов (Brookhaven National Laboratory, США)

Михаил Когаловский (Институт проблем рынка РАН)

Владимир Кореньков (ОИЯИ, Дубна)

Ефим Кудашев (Институт космических исследований РАН)

Сергей О. Кузнецов (НИУ ВШЭ)

Сергей Д. Кузнецов (ИСП РАН)

Дмитрий Ландэ (Институт проблем регистрации информации НАН Украины)

Владимир Литвин (Evogh Inc., Калифорния, США)

Джузеппе Лонго (Неапольский университет)

Наталья Лукашевич (НИВЦ МГУ им. М. В. Ломоносова)

Иван Лукович (Нови-Садский университет)

Яннис Манолопоулос (School of Informatics of the Aristotle University of Thessaloniki)

Александр Марчук (A.P. Ershov Institute of Informatics Systems SB RAS)

Пётр Мировски (Google Deep Mind, London)

Арчил Майсурадзе (ВМК МГУ)

Олег Малков (Институт астрономии РАН)

Игорь Некрестьянов (Verizon Corporation, США)

Борис Новиков (Санкт-Петербургский государственный университет)

Геннадий Ососков (Объединённый институт ядерных исследований)

Дмитрий Палей (Ярославский государственный университет)

Стелиос Папаризос (Google, Сан-Франциско)

Милан Петкович (Technical University of Eindhoven)

Николай Подколодный (ИЦиГ СО РАН)

Алексей Позаненко (ИКИ РАН, Москва)

Борис Позин (ЕС-Лизинг)

Ярослав Покорны (Карлов Университет, Прага)

Наталия Пономарева (Научный центр неврологии РАМН)

Андреас Раубер (Vienna TU)

Александр Рогов (Петрозаводский государственный университет)

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

Николай Скворцов (ФИЦ ИУ РАН, Москва)

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

Леонид Соколинский (Южно-Уральский государственный университет, Челябинск)

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

Сергей Ступников (ФИЦ ИУ РАН, Москва)

Александр Сычев (Воронежский государственный университет)

Бернард Тальхайм (Кильский университет, Германия)

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)

Сопредседатели программного комитета диссертационного семинара

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

Илья Соченков (ФИЦ ИУ РАН)

XVIII International Conference DAMDID / RCDL'2016

Data Analytics and Management in Data Intensive Domains

October 11-14, 2016, Ershovo, Moscow, Russia

http://damdid2016.frccsc.ru/

Coordinating Committee of DAMDID/RCDL Conferences

Co-Chairs

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

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, SKB Kontur, Russia

Vasily Bunakov – Science and Technology Facilities Council, Harwell, Oxfordshire, UK

Alexander Elizarov – Kazan (Volga Region) Federal University, Russia

Alexander Fazliev – Institute of Atmospheric Optics, RAS, Siberian Branch, Russia

Alexei Klimentov - Brookhaven National Laboratory, USA

Mikhail Kogalovsky – Market Economy Institute, RAS, 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

Archil Maysuradze – Moscow State University, Russia

Oleg Malkov – Institute of Astronomy, RAS, Russia

Alexander Marchuk – Institute of Informatics Systems, RAS, Siberian Branch, 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

Vladimir Smirnov – Yaroslavl State University, Russia

Sergey Stupnikov – Federal Research Center "Computer Science and Control" of RAS

Konstantin Vorontsov – 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'2016 Conference

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

Organizing Committee

Co-chairs

Boris Pozin, MIEM NRU HSE

Victor Zakharov, FRC CSC RAS

Alexander Nevzorov, NRNU MEPhI

Members of Organizing Committee

Dmitry Briukhov, FRC CSC RAS, web site support

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

Nkolay Skvortsov, FRC CSC RAS, CMT management, technical editing of conference materials

Evgenia Dudareva, FRC CSC RAS, financial issues

Anna Nadezhdina, FRC CSC RAS, treasurer

Yulia Trusova, FRC CSC RAS, visa support

Tatjana Dudareva, Protey Travel, local arrangement chair

Sergey Vereshchagin, INASAN, organizing committee relationships with Ershovo Holiday Center

Protey Travel, transportation of the conference participants

MIEM NRU HSE, Ershovo Holiday Center, technical support, including PC, projectors, Wi-Fi in conference rooms and in lodging

MEPhI, synchronous translation

Program committee

Co-chairs

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

Sergei O. Kuznetsov, National Research University Higher School of Economics

Yannis Manolopoulos, Aristotle University of Thessaloniki, Greece

Members of the Program Committee

Karl Aberer (EPFL, Lausanne, Switzerland)

Plamen Angelov (Lancaster University, UK)

Arkady Avramenko (Pushchino Observatory)

Ladjel Bellatreche (LIAS/ISAE-ENSMA, Poitiers, France)

Pavel Braslavski (Ural Federal University, Yekaterinburg)

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

Panos Chrysanthis (University of Pittsburgh)

Paul Cohen (School of Information, University of Arizona)

Ernesto Damiani (Khalifa University, Abu Dhabi)

Yuri Demchenko (University of Amsterdam)

Boris Dobrov (Research Computing Center of MSU)

Shlomi Dolev (DCS, Ben-Gurion University of the Negev, Beer-Sheva, Israel)

Alexander Elizarov (Kazan Federal University)

Alexander Fazliev (Institute of Atmospheric Optics, SB RAS)

Olga Gorchinskaya (FORS, Moscow)

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

Maxim Gubin (Google Inc.)

Ralf Hofestadt (University of Bielefeld)

Leonid Kalinichenko (FRC CSC RAS, Moscow)

George Karypis (University of Minnesota, Minneapolis)

Nadezhda Kiselyova (IMET RAS)

Alexei Klimentov (Brookhaven National Laboratory, USA)

Mikhail Kogalovsky (Market Economy Institute, RAS)

Vladimir Korenkov (Joint Institute for Nuclear Research, Dubna)

Efim Kudashev (Space Research Institute, RAS)

Dmitry Lande (Institute for Information Recording, NASU)

Vladimir Litvine (Evogh Inc., California, USA)

Giuseppe Longo (Naples)

Natalia Loukachevitch (Lomonosov MSU)

Ivan Lukovic (University of Novi Sad)

Oleg Malkov (Institute of Astronomy, RAS)

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

Alexander Marchuk (A.P. Ershov Institute of Informatics Systems SB RAS)

Archil Maysuradze (MSU)

Piotr Mirowski (Google Deep Mind, London)

Igor Nekrestyanov (Verizon Corporation, USA)

Boris Novikov (Saint Petersburg State University)

Gennady Ososkov (Joint Institute for Nuclear Research)

Dmitry Paley (Yaroslav State University)

Stelios Paparizos (Google, San Francisco)

Milan Petkovic (Technical University of Eindhoven)

Nikolay Podkolodny (Institute of Cytology and Genetics SB RAS)

Jaroslav Pokorny (Karlov University in Prague)

Natalia Ponomareva (Scientific Center of Neurology of RAMS)

Alexey Pozanenko (Space Research Institute, RAS)

Boris Pozin (EC-Leasing)

Andreas Rauber (Vienna TU)

Alexander Rogov (Petrozavodsk State University)

Timos Sellis (RMIT, Australia)

Vladimir Serebryakov (Computing Centre of RAS)

Nikolay Skvortsov (FRC CSC RAS)

Vladimir Smirnov (Yaroslavl State University)
Leonid Sokolinskiy (South Ural State University)
Valery Sokolov (Yaroslavl State University)
Sergey Stupnikov (FRC CSC RAS)
Alexander Sychev (Voronezh State University)
Bernhard Thalheim (University of Kiel)
Dmitry Tsarkov (Manchester University)
Alexey Ushakov (University of California, Santa Barbara, USA)
Natalia Vassilieva (Hewlett-Packard)
Konstantin Vorontsov (VMK Faculty, MSU)
Alexey Vovchenko (FRC CSC RAS, Moscow)
Domagoj Vrgoc (Center for Semantic Web Research)
Peter Wittenburg (MPI for Psycholinguistics)
Viacheslav Wolfengagen (MEPhI)
Yury Zagorulko (Institute of Informatics Systems, SB RAS)
Victor Zakharov (FRC CSC RAS)
Oleg Zhizhimov (Institute of computing technologies of SB RAS)
Sergey Znamensky (Institute of Program Systems, RAS)

PhD Workshop Co-Chairs

Ladjel Bellatreche, LIAS/ISAE-ENSMA, Poitier, FRANCE
Ilia Sochenkov, FRC CSC RAS

Содержание / Contents

Предисловие

В 2016 году Международная конференция «Аналитика и управление данными в областях с интенсивным использованием данных» (“Data Analytics and Management in Data Intensive Domains” – DAMDID/RCDL’2016) проводилась с 11 по 14 октября в доме отдыха Ершово (Московская область, Одинцовский район).

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). Подходы к анализу данных и управлению данными, развиваемые в конкретных ОИИД X-информатики (таких как X=астро, био, гео, нейро, медицина, физика, химия, и пр.), социальных наук, а также различных ОИИД информатики, промышленности, новых технологий, финансов и бизнеса составляют предметную область конференции.

Программа конференции 2016 года отражает наряду с традиционной для управления данными тематикой движение в направлении науки о данных (data science) и аналитики с интенсивным использованием данных. Три приглашенных пленарных доклада конференции акцентируют ключевые проблемы развития методов и средств аналитики и управления данными в ОИИД. В докладе Андрея Ржецкого (профессора генетической медицины Чикагского университета), открывающем конференцию, рассматриваются подходы к изучению механизмов рака, развиваемые консорциумом UChicago, работающим под руководством Андрея Ржецкого по программе DARPA Big Mechanism (http://www.darpa.mil/program/big-mechanism). Эти подходы сфокусированы на развитии современных методов анализа данных в ОИИД, включая когнитивные методы, основанные на понимании естественного языка, подобные методам системы Ватсон IBM, с упором на формирование гипотез на основе обнаруживаемых в текстах причинных отношений, моделирование механизмов рака для автоматического предсказания терапевтических решений, применение методов организации экспериментов на основе их роботизации. Второй день открывает доклад Софии Ананиаду (директора Национального исследовательского центра Соединенного королевства по извлечению информации из текстов (NaCTeM) при Манчестерском университете, входящего составной частью в консорциум UChicago, финансируемого DARPA), в котором рассматриваются проблемы автоматической реконструкции траекторных моделей как результата обнаружения отношений между понятиями различной природы в текстах. Наконец, программа третьего дня начинается докладом Димитриоса Тзовараса (профессора и руководителя Института информационных технологий Центра исследований и технологий Hellas (Эллада) в Салониках), в котором дан аналитический обзор Европейской стратегии в области исследовательских инфраструктур.

Программный комитет конференции рассмотрел 57 заявок. Из них: 27 приняты как полные статьи, 16 – как краткие, 3 – как постеры, 2 – как демо, 9 – отклонены. 43 доклада (полные и краткие) представлены в 13 сессиях, таких как семантическое моделирование в DID, методы анализа данных, управление знаниями, системы управления обучением, семантический поиск и навигация, анализ паттернов в рекомендательных системах, исследовательские инфраструктуры (в астрономии, астрофизике, материаловедении, для мониторинга Земли), извлечение информации из текстов, интеграция и совместное использование данных, системы анализа текстов. Большинство докладов посвящены результатам исследований, выполняемых в исследовательских организациях, расположенных в различных местах на территории России, включая: Дубну, Екатеринбург, Звенигород, Иркутск, Казань, Москву, Новосибирск, Обнинск, Пущино, Томск, Тулу, Санкт Петербург, Ярославль.

Кроме того, в состав конференции включены следующие ассоциированные мероприятия: тьюториал, подготовленный профессором Вереной Кантере и ее аспирантом Максимом Филатовым в Женевском университете и содержащий анализ эффективности средств поддержки аналитики данных в многомашинных средах; диссертационный семинар, содержащий 5 докладов; а также открытый симпозиум «Здоровьесбережение на основе интенсивного использования данных», включающий три круглых стола, на которых обсуждаются бизнес-модели, мобильные технологии, вопросы диагностики в персонализированной медицине, а также две сессии, тексты докладов на которых, посвященные рассмотрению роли информатики в Здоровьесбережении, включены в сборник трудов конференции. Симпозиум подготовлен по инициативе члена-корреспондента РАН Андрея В. Лисицы (Научно-исследовательский институт биомедицинской химии (ИБМХ)).

Председатели Программного и Организационного комитетов конференции выражают благодарность Николаю Скворцову за взаимодействие при помощи systems CMT с авторами присланных работ и с членами ПК – рецензентами докладов, а также за подготовку верстки сборника трудов конференции в процессе издания его печатной версии. 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’2016.

Дом отдыха Ершово, в котором проводилась конференция, расположен в уникальном историческом уголке Подмосковья в усадьбе династии Олсуфьевых. Усадьба Ершово находится в 50 км от Москвы в окрестностях Звенигорода, одном из красивейших мест Подмосковья, где неповторимая природа, живописный ландшафт и чистый воздух создают особую атмосферу для проведения научных форумов. Обладая современной инфраструктурой, удобной логистикой (порядка одного часа до Москвы по железной и автодорогам), историческими и природными достопримечательностями, Ершово способствовало проведению очередной научной конференции DAMDID/RCDL на высоком уровне.

Председатели Организационного комитета и Программного комитета конференции выражают благодарность авторам поданных на конференцию заявок, а также Российскому Фонду Фундаментальных Исследований и Национальному исследовательскому ядерному университету МИФИ за финансовую поддержку конференции.

Сопредседатели
Программного комитета

Калиниченко Леонид Андреевич
(ФИЦ ИУ РАН)
Кузнецов Сергей Олегович
(НИУ ВШЭ)
Манолопулос Янис
(Университет Аристотеля, Салоники)

Сопредседатель
Организационного комитета

Захаров Виктор Николаевич
(ФИЦ ИУ РАН)

Preface

In 2016 the International Conference “Data Analytics and Management in Data Intensive Domains” (DAMDID/RCDL’2016) was held on October 11 – 14 in the Holiday Center, Ershovo (Moscow region).

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, medicine, neuro, physics, 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’2016 conference alongside with the traditional data management topics reflects a rapid move into the direction of data science and data intensive analytics. Three invited plenary talks of the conference emphasize the key problems of methods and facilities of data analytics and management in DID. In the keynote of Andrey Rzhetsky (Professor of genetic medicine of the Chicago University) that opens the conference, the approaches for studying the mechanisms of cancer are presented, which are being developed by the UChicago consortium in which Andrey Rzhetsky acts as the PI according to the DARPA Big Mechanism program ((http://www.darpa.mil/program/big-mechanism). These approaches are focused on the development of the contemporary methods of data analysis in DID including cognitive methods based on the understanding of the natural language resembling those of IBM Watson center, but focused on the hypothesis formation based on the causal relationships detected in the texts, cancer mechanism modeling to automatically predict therapeutic clues, application of robot scientist methods for the organization of experiments. The invited talk of Sophia Ananiadou (Director of the National Centre for Text Mining (NaCTeM) at the Manchester University) opens the second day of the conference. NaCTeM is included as a part of the UChicago consortium supported by DARPA. In this talk the approaches for automatic reconstruction of the pathway models are considered. These approaches are based on the discovery of the relationships of various kinds between the concepts of arbitrary nature in the texts. Finally, the program of the third day is opened with the keynote of Dimitrios Tzovaras (Director at the Information Technologies Institute of the Centre for Research and Technology Hellas in Thessaloniki) in which the analytical survey of the European strategy in the area of research infrastructures is presented.

The conference Program Committee has reviewed 57 submissions and accepted of them 27 as full papers, 16 as short papers, 3 as posters, 2 as demos, whereas 9 submissions were rejected. According to the conference program, these 43 oral presentations (of the full and short papers) are structured into 13 sessions including Semantic Modeling in DID, Data Analysis Methods, Knowledge Management, Learning Management, Semantic Search and Navigation, Pattern Analysis in Recommender Systems, Research Data Infrastructures (in Astronomy, Astrophysics, Material Sciences, Earth Monitoring), Data Extraction from Texts, Data Integration and Sharing, Text Analysis Systems. 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 Dubna, Ekaterinburg, Irkutsk, Kazan, Moscow, Novosibirsk, Obninsk, Puschino, Tomsk, Tula, Saint Petersburg, Yaroslavl, Zvenigorod,

Besides the above, the conference contains also the following satellite events: a tutorial “Data Analytics in Multi-Engine Environments” prepared by Dr. Verena Kantere and Ph.D. student Maxim Filatov from the University of Geneve; a Ph.D. Workshop containing 5 presentations; an open workshop “Data Intensive Healthcare”, including three round tables with the discussions of business models, mobile technologies, diagnostics and coaching in the personalized medicine, as

well as two sessions entitled as “Informatics in Healthcare” (the papers reflecting the presentations at these sessions are included into the conference proceedings). This workshop is prepared according to the initiative of Andrey Lisitsa (corresponding member of RAS, Research Institute of the Biomedical Chemistry, Moscow).

The chairs of the Program Committee and Organizing Committee of DAMDID/RCDL’2016 express their gratitude to Nikolay Skvortsov for the effective interactions by the CMT system with the authors of submissions and with the PC members reviewing the submissions, as well as for preparing a layout of the conference proceedings during the process of its publishing. 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 Ershovo Holiday Center in which the conference was held is located in the historic Ershovo mansion placed at the unique historical corner in the vicinity of Moscow, in the neighborhood of Zvenigorod. Providing an up-to-date infrastructure, convenient logistics (50 km from Moscow, about one hour to Moscow by railway or highway), historical and natural attractions, the venue contributed to the organization of the DAMDID/RCDL conference on a high standard.

The chairs of the Organizing Committee and Program Committee of DAMDID/RCDL’2016 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 for the financial support to the Conference.

Co-chairs of the Program committee

Leonid A. Kalinichenko
(FRC CSC RAS)
Sergey O. Kuznetsov
(NRU HSE)
Yannis Manolopoulos
(Aristotle University, Thessaloniki)

Co-chair of the Organizing committee

Victor N. Zakharov
(FRC CSC RAS)

Ключевой доклад 1

Keynote Talk 1

The Big Mechanism Program: Changing How Science Is Done

Andrey Rzhetsky
University of Chicago, 900 East 57th Street, Chicago, IL 60637, USA

andrey.rzhetsky@uchicago.edu

Abstract

The talk will describe details of actively evolving research conducted by the UChicago consortium of the Big Mechanism program, funded by the US DARPA agency. The consortium's work focuses on: (1) probabilistic reasoning across cancer claims culled from literature which uses custom-designed ontologies; (2) the computational modelling of cancer mechanisms and pathways to automatically predict therapeutic clues; (3) automated hypothesis generation to strategically extend this knowledge, and; (4) developing a 'Robot Scientist' that performs experiments to test hypotheses probabilistically, then feeding those results back to the system.

1 Introduction

DARPA is funding the Big Mechanism program (http://www.darpa.mil/program/big-mechanism) in order to study large, explanatory models of complicated systems in which interactions have important causal effects. The program's aim is to develop technology used to read research abstracts and papers and extract pieces of causal mechanisms, assemble these pieces into more complete causal models, and reason over these models to produce explanations. The program's domain is cancer biology, with an emphasis on signalling pathways; this is just one example of causal, explanatory models, that we are hoping will be extensible across multiple domains, similar to what IBM Watson's team [1] is attempting presently.

2 The overall structure of the Big Mechanism program

The program is currently organized into three consortia, all of which take different views of causal models, different reading technologies, and different use cases.

The largest consortium, called FRIES, includes groups at CMU, SRI, University of Arizona, Oregon Health Sciences University, and others. FRIES's main focus is to explain signalling pathway behaviours. For instance, why is the expression of a gene ephemeral? Technologically, FRIES focuses on information extraction over deep reading, simulation, and even FPGA acceleration of systems biology simulators.

The second consortium ("UChicago"), in which the author of this keynote acts as the PI, is composed of researchers at the University of Chicago, the United Kingdom's National Center for Text Mining at the University of Manchester, along with participants from the Brunel University in London, all of whom collaborate on developing robotic platforms for experiment design and analysis.

The third consortium, called CURE, consists of two groups from Harvard Medical School, IHMC in Florida, and SIFT. Their focus is on deep reading, fine-grained modeling, and simulation of cell signaling's underlying biochemistry.

This talk will provide an overview of the objectives and results related mostly to the work of the second consortium.

3 UChicago consortium

As the project is ongoing and far from completion, we will cover the ideas that led the consortium to our current system design, our biological and medical motivations, and preliminary results.

Motivation: Today, cancer-related text mining is performed in linear pipelines (named entity recognition to event extraction) without explicitly estimating statement uncertainty or importance relative to a total model of cancer. Moreover, reading is divorced from reasoning and experimentation. Probabilistic reasoning is rarely used. Similarly, the Robot Scientist approach currently uses non-probabilistic logic and is disconnected from text mining and not applied to medicine. In addition, a wealth of panomics data is increasingly available, but existing methods treat each event independently and disregard prior knowledge.

Fundamental medical problem: We do not fully understand how to stop cancer cells from growing faster than normal tissue, and spreading throughout the body (metastasizing). Death from cancer typically occurs when uncontrolled growth occurs in a place where it cannot be surgically removed. Most traditional anti-cancer drugs are highly toxic to patients. As a result, single drug treatment is generally undesirable for the following reasons: (1) It is generic and not targeted to the patient and their cancer's genotype(s); (2) Intervention is

Proceedings of the XVIII International Conference «Data Analytics and Management in Data Intensive Domains» (DAMDID/RCDL'2016), Ershovo, Russia, October 11 - 14, 2016

required at multiple points along a cancer pathway, and; (3) Cancer evolves resistance. The Holy Grail of cancer therapy is to find highly potent, non-toxic drug combinations that are tailored to individual patients, and linked to the readout of gene and protein expression from their specific cancer(s).

The system developed by the consortium incorporates three components, called Reading, Assembly, and Explanation (see Figure 1). These components integrate machine reading with probabilistic modelling, the design of custom-made ontologies, and automated experiments conducted by the Robot Scientist (a robot that is driven by experiment-designing and planning programs). For quality control and benchmarking, an independent set of experiments is conducted by humans.

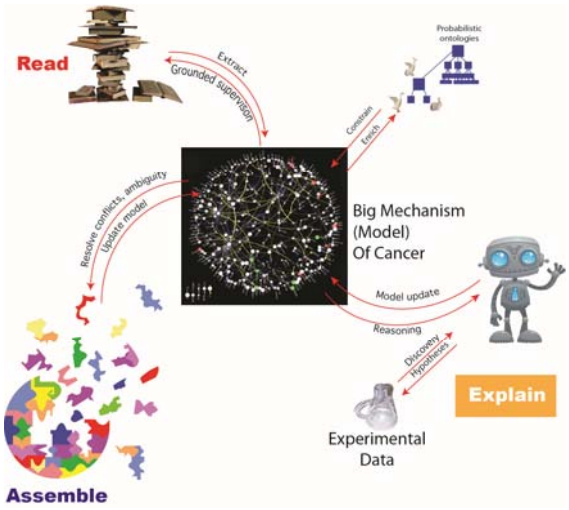


Figure 1 The integrated system, see references [2,3] for related prior work contributing to the components of the system

To illustrate how all these components come together, the talk will present a use case: Automated, optimal drug combination prediction for achieving activation or silencing of target gene(s) in a breast cancer cell line. In our initial setup, we are using a text-mined network of about three hundred genes and proteins, containing parts of networks in use cases 1 and 2. In the first pass, we focused on activating the estrogen receptor gene (ESR1) in a triple-negative breast cancer cell line by administering a cocktail of two or more FDA-approved drugs.

The motivation for the use case is to practically apply growing (through machine reading and experimental validation) model of cellular machinery to manipulate the state of the cancer cell, achieving silencing or activation of target genes/proteins in the absence of drugs specifically targeting these molecules. If successful, computationally-derived drug cocktails could at least partially reduce the need to develop new drugs, easing the economic burden of discovering and testing new medications. (Each new FDA-approved drug has an estimated price tag of somewhere between 100 million and 1 billion US dollars.)

The system generates hypotheses of the form "cocktail of drugs X1, …, Xn activates gene ESR1" and each hypothesis is tested experimentally in a triple-negative breast cancer cell line. Either human biologists or the Robot Scientist carry out these experiments.

4 "UChicago" team

Reading (NLP and text-mining; ontologies, corpus-dependent and unsupervised information extraction, logic): Sophia Ananiadou, Junichi Tsujii, Larisa Soldatova, Hoifung Poon, Andrey Rzhetsky, Robert Stevens, James Evans.

Assembling (Models of quality of science, quality of extraction, consistency, statement provenance, Markov Logic, crowdsourcing): Jacob Foster, James Evans, Hoifung Poon, Andrey Rzhetsky.

Explaining (Markov Logic, graphical models, consistency models, kinetic/dynamic consistency models): Hoifung Poon, Jacob Foster, James Evans, Ishanu Chattopadhyay, Andrey Rzhetsky.

AI and Robotics: Hoifung Poon, Kevin P. White, Ross D. King. Cancer-specific, wet-lab experiments: Ross D. King, Kevin P. White.

In prior work, Ross D. King's laboratory has developed two Robot Scientists, "Adam" and "Eve", which are among the most advanced existing laboratory automation systems.

5 Conclusion

The approach chosen by the team relies on the assimilation of massive, pre-existing literature (similar to IBM Watson) combined with iterative model updating based on empirical data and newly designed experiments (unlike IBM Watson). The project's general methodology is not domain-specific, so it is theoretically extensible across scientific domains.

References

[1] Best, J. IBM Watson: The Inside Story Of How The Jeopardy-Winning Supercomputer Was Born, And What It Wants To Do Next - Feature - TechRepublic. TechRepublic. N.p., 2015. Web. 13 May 2015.

[2] Evans J, Rzhetsky A. Machine science. Science. Jul 23 2010;329(5990):399-400.

[3] King RD, Rowland J, Oliver SG, Young M, Aubrey W, Byrne E, Liakata M, Markham M, Pir P, Soldatova, LN, Sparkes A, Whelan KE, Clare C. The Automation of Science. Science. 2010. 324, 85-89.

Семантическое моделирование в DID

Semantic modeling in DID

Metadata for nanoscience experiments

© Vasily Bunakov © Tom Griffin ©Brian Matthews
Science and Technology Facilities Council,
Harwell, Oxfordshire, UK
vasily.bunakov@stfc.ac.uk tom.griffin@stfc.ac.uk brian.matthews@stfc.ac.uk
© Stefano Cozzini
Instituto Officina dei Materiali
Trieste, Italy
cozzini@iom.cnr.it

Abstract

Metadata is a key aspect of data management. This paper describes the work of NFFA project on the design of a metadata standard for nanoscience community. The methodology and the resulting high-level metadata model are presented. The paper explains and illustrates the principles of metadata design for data-intensive research. This is value to data management practitioners in all branches of research and technology that imply a so-called "visitor science" model where multiple researchers apply for a share of a certain resource on large facilities (instruments).

1 Introduction

The Nanostructures Foundries and Fine Analysis (NFFA-Europe) project www.nffa.eu brings together European nanoscience research laboratories that aim to provide researchers with seamless access to equipment and computation. This will support a single entry point for research proposals supported by the project, and a common platform to support the access and integration of the resulting experimental data. Both physical and computational experiments are in scope, with a vision that they complement each other and can be mixed in the same identifiable piece of research.

The project requires setting up the IT infrastructure for managing research proposals and substantial amounts of data resulted from physical and computational experiments. A common metadata model that supports different stages of the nanoscience research lifecycle is essential to unified researchers' experience across locations, and also for the design and operation of IT infrastructure components.

Metadata design is a part of a joint research activity within NFFA that takes empirical input from the project participants, also takes into account state-of-the art standards and practices. Metadata design is an incremental effort of the project; this work presents the first stage resulting in a high-level metadata model that is agnostic to the actual data management situation in participating organizations yet is able to capture significant features of nanoscience physical and computational experiments.

Proceedings of the XVIII International Conference «Data Analytics and Management in Data Intensive Domains» (DAMDID/RCDL'2016), Ershovo, Russia, October 11 - 14, 2016

2 Approach and methodology

2.1 General approach

The major purpose of any metadata is satisfying information needs of a certain community. "Community" should be understood in broad terms and includes machine agents, to ensure human-to-human, human-to-machine and machine-to-machine interoperability.

The information needs may be generic (common with other communities) or specific for a particular community. From the project perspective, the information needs should be expressed as clearly formulated Use Cases for the existing or proposed information and data management systems (IT platforms). A good metadata design should take into account user requirements and IT architecture, and in turn should feed considerations for the IT architecture.

The IT architecture, the use cases and practices, and the metadata design can be considered pillars of *enterprise architecture* that includes both technological and organizational aspects of a loosely coupled virtual enterprise that the NFFA project is going to deliver for the European nanoscience community.

The main purpose of metadata design effort in NFFA project can be formulated then as giving the adequate support for that widely defined enterprise architecture for nanoscience. This has an implication of metadata design from "first principles", i.e. by pondering over existing best practices of information management, use cases for nanoscience and information technology opportunities (and limitations) rather than adopting any existent metadata standard.

2.2 Top-down input: relevant information management frameworks

The case for metadata collection and use can be specific to nanoscience, yet there are general information needs that are typical for a wide variety of users and that have been developed in other branches of science and information management.

One of the mature information design frameworks is Functional Requirements for Bibliographic Records [2]

that considers four basic information needs (user tasks) in regards to information: “Find”, “Identify”, “Select” and “Obtain”. The ultimate goal is of course getting the information resource, yet between searching for it and obtaining it, the resource should be identified as the one being sought, and selected as being useful for the user [1]. Each task may involve certain subtasks, e.g. selection may require checks on the resource context and on its relevance to the actual user’s needs.

Another mature information design framework of relevance is the Reference Model for an Open Archival Information System [3], a popular functional model for long-term digital preservation. If expressed in terms of information practitioner needs (user tasks) similarly to FRBR, the OAIS basically deals with three categories of them: “Ingest (into the archive)”, “Manage (within the archive)” and “Disseminate (from archive)”. Each of these tasks may be complex and involve a number of interrelated subtasks, e.g. managing information in the archive may imply provenance and integrity checks, managing access to information, and administration / reporting.

Overall, the OAIS framework should be able to provide a good coverage of what NFFA needs to consider for sensible data collection, archiving and provision towards the end users (researchers in nanoscience), and the FRBR framework should be able to cover the end user needs for information retrieval. The respective areas of coverage and user categories relevant to NFFA are illustrated by the following table:

Table 1 Information management frameworks and their coverage of NFFA scope.

Framework (a source of best practices)	**OAIS**	**FRBR**
General use case	Data collection, management and dissemination	Data retrieval
User categories	Data archives administrators IT specialists	End users (nanoscience researchers)
Information needs (user tasks)	Ingest data Manage data Disseminate data	Find data Identify data Select data Obtain data

Being general in nature, OAIS and FRBR are still able to provide good recommendations for NFFA practices of information and data management. In particular, OAIS emphasizes the need of having a clear agreement between the data producer and the archive, and a clearly defined format for data exchange between them – so called Submission Information Package, whilst FRBR emphasizes the importance of having a clear identity for data assets.

2.3 Bottom-up input: questionnaire responses and common vocabulary

A questionnaire was used to collect the NFFA partners’ responses about their data management practices and most popular data management solutions. The questionnaire inquired on the following aspects of data management in nano-facilities:

- Intensity of experiments and of resulting data flow
- Popular data formats
- Data catalogue software
- Data catalogue openness
- Data management policy
- Metadata standards for data catalogue
- Persistent identifiers for data
- User management platform
- Popular third-party databases and information systems

In total, seventeen responses out of the 20 project partners were received and reviewed. They showed very different levels of data management maturity. From the responses, the following priorities the metadata design were identified:

- One experiment to many samples and one sample to many data files relationships should be supported.
- A common set of metadata fields for data discoverability should be agreed upon, possibly based on an existing popular standards or recommendation for data discovery.
- User roles with different permissions for access to metadata should be developed. This means the metadata model will need to represent users as well as data.
- It is reasonable to develop a common data management policy for NFFA, or a set of policies with different flavours of access to data.
- Having links to external reference databases is valuable to ensure the high quality of metadata yet this will mean additional effort so should be de-scoped from the initial design of metadata.

In addition to the questionnaire where responses were collected from research offices or relevant research programme representatives, a common vocabulary of terms and definitions relevant to nanoscience data management was compiled and then refined by the IT teams of participating NFFA organizations ([5]). The vocabulary contains about twenty commonly agreed terms with definitions; it serves as a basis for the design of information entities (groups of metadata elements) and contributes to the earlier mentioned NFFA “virtual enterprise” architecture.

A particularly important use case to be supported by the metadata model should be the situation when the same researcher (or a research group) applies for experimental time on more than one facility – as the nature of experiment may require this – yet the researcher wants a seamless experience across nanoscience facilities, with a single entry point for data management.

Another conclusion based on responses to the questionnaire is that computational experiments in nanoscience become common and can be mixed up with physical experiments, so there should not be an artificial division between the two.

2.4 Side input: IT architecture considerations

As an additional consideration for principal metadata design, we used the draft NFFA Data System Architecture that defines the outline design of the NFFA portal, which considered the generic use case of the same user performing a measurement on two different facilities. Generic use cases when one user wants to access data produced by another user, or wants to release data into the public domain are currently not being considered. These may be considered in future, so should be taken into account within an extensible metadata design.

The draft architecture suggests that data should be harvested from individual facilities in a suitable "packaged" format, with METS [6] as a potential candidate as it supports the provision of descriptive, administrative, structural and file metadata. For the descriptive part of metadata, the purpose of having the data assets discoverable is emphasized in the draft architecture. For the administrative metadata, the importance of intellectual property information and information about the data source (provenance) is emphasized. For the structural metadata, having the information about the organization, perhaps structured in a hierarchical way, is suggested. For the file metadata, having the list of files that constitute a digital object (data asset) and having pointers to external metadata files are deemed most important.

After considering the draft architecture, the conclusion was that we could take METS as "the role model" metadata standard for data packaging that corresponds to a specific entity in the NFFA generic metadata model – Data Asset. As to particular elements of metadata suggested by the IT architecture draft, the fields for capturing intellectual property information and provenance are easily most important ones as they affect the data assets reusability that should be one of the important outcomes of the NFFA project.

3 Implementation

3.1 Metadata groups and elements

The top-down, bottom up and side requirements resulted in the basic structure of the proposed metadata model that is illustrated by Figure 1.

The suggested metadata elements are presented as a matrix in Table 2 to make explicit the coverage of identified information entities (common vocabulary terms) and of earlier identified information needs categories of them, see Section 2.2).

Certain elements are in common with the Core Scientific Metadata Model ([4]) already in use in some of the facilities. Mandatory and optional metadata fields (attributes) for each element were defined and shared amongst project participants for further discussion ([5]).

3.2 Entity-relationship diagram

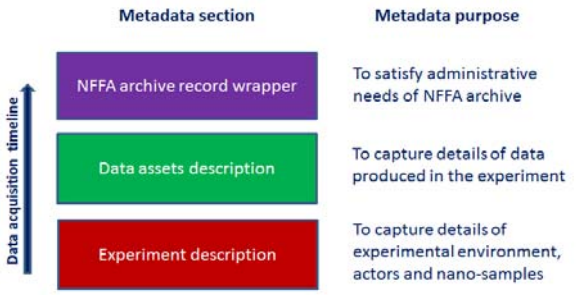


Figure 1 Metadata groups of elements and their purpose.

Table 2 Metadata elements and information needs coverage.

Information entity	Ingest data	Manage data (within NFFA portal)	Disseminate data	Find data	Identify data	Obtain data
Research User			Y	Y	Y	Y
Instrument Scientist	Y	Y				
Project			Y	Y	Y	Y
Proposal	Y	Y				
Facility	Y	Y	Y	Y	Y	Y
Instrument			Y	Y	Y	
Experiment			Y	Y	Y	
Sample			Y	Y	Y	
Data Asset	Y	Y	Y	Y	Y	Y
Raw Data	Y	Y	Y	Y	Y	Y
Analysed Data	Y	Y	Y	Y	Y	Y
Data Analysis	Y	Y			Y	
Data Analysis Software	Y	Y			Y	
Data Archive	Y	Y				Y
Data Manager	Y	Y				Y
Data Policy	Y	Y				
NFFA Portal		Y		Y		

As a basis for further, more detailed metadata design and as a contribution to the IT architecture design, the Entity-Relationship diagram presented by Figure 2 has been agreed.

3.3 Metadata operational recommendations

The metadata elements suggested are not all we need for having a successful metadata framework in NFFA. In

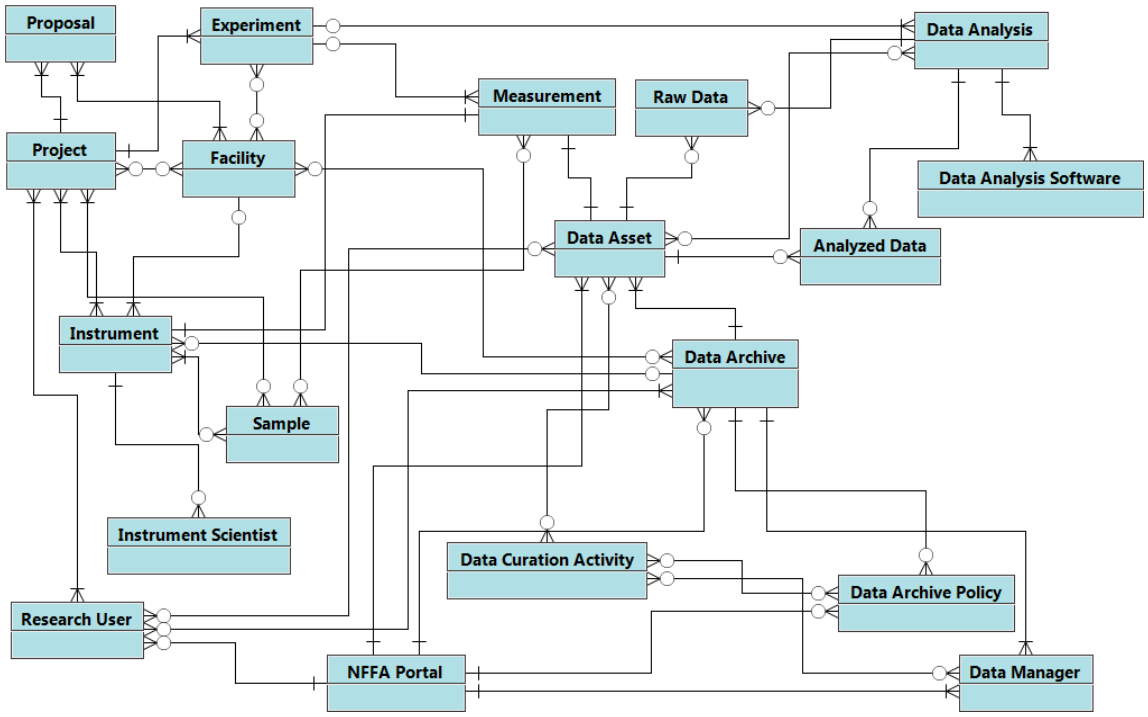


Figure 2 Entity-Relationship diagram for NFFA high-level metadata model.

addition, there should be established metadata management practices, ideally assisted by clear recommendations for NFFA partner organizations of how to assign and curate metadata.

For example, there are choices of how you aggregate data: let us say all data files for all samples measured in a particular Experiment can be assembled in one package, and then the package is given common descriptions like Facility name, research User name, Data Policy etc. However, this may not suit actual data management practices or policies of certain Facilities, e.g. they may want to make a Sample rather than an Experiment a focal point of their metadata descriptions.

These operational aspects of NFFA metadata implementation will require further engagement and discussions with data practitioners in NFFA.

4 Conclusion

The NFFA metadata development so far has produced an agreed common approach with its mapping to the existing metadata frameworks and best practices. It has defined a common vocabulary, the provisional list of mandatory and optional attributes, and the ER diagram that can be used both in metadata design and in IT architecture design. The high-level metadata model will be further refined through project work in NFFA and through discussions in the wider nanoscience community. Also the state-of-the-art metadata development for nanoscience that may cover specific entities in our generic metadata model, e.g. CODATA UDS [7] for Sample, should be looked into in more detail, to see the opportunities for mutual mapping and cross-walks between different metadata models.

Acknowledgements

This work is supported by Horizon 2020 NFFA-Europe project www.nffa.eu under grant agreement no. 654360.

References

[1] Philip Hider. Information resource description: Creating and managing metadata. Facet Publishing, 2012.

[2] Functional Requirements for Bibliographic Records (FRBR). Final Report. http://archive.ifla.org/archive/VII/s13/frbr/ Retrieved 20 May 2016.

[3] Reference Model for an Open Archival Information System (OAIS), Recommended Practice, CCSDS 650.0-M-2 (Magenta Book). Issue 2, June 2012. http://public.ccsds.org/publications/archive/650x0m2.pdf Retrieved 20 May 2016.

[4] The Core Scientific Metadata Model (CSMD). https://icatproject.org/user-documentation/csmd/ Retrieved 20 May 2016.

[5] Draft metadata standard for nanoscience data. NFFA project deliverable D11.2. February 2016.

[6] METS: Metadata Encoding and Transmission Standard. http://www.loc.gov/standards/mets/

[7] COADATA UDS: Uniform Description System for Materials on the Nanoscale http://www.codata.org/uploads/Uniform_Description_System_Nanomaterials-Published-v01-15-02-01.pdf

Концептуальное моделирования предметных областей с интенсивным использованием данных

© Н. А. Скворцов © Л. А. Калиниченко © Д. Ю. Ковалев

Институт проблем информатики ФИЦ ИУ РАН, Москва

nskv@mail.ru leonidandk@gmail.com dm.kovalev@gmail.com

Аннотация

Исследования в различных предметных областях, особенно в направлениях естественных наук, связаны сегодня с обработкой больших объёмов данных наблюдений, экспериментов и моделирования. При организации исследований с интенсивным использованием данных целесообразно определять спецификации предметных областей, включающие определения понятий предметных областей средствами онтологий и абстрактное представление данных об объектах предметных областей и их поведении средствами концептуальных схем, разделяемых и поддерживаемых работающими в этих предметных областях сообществами. Исследовательские инфраструктуры опираются на спецификации предметных областей и предоставляют реализации методов, применимых над такими спецификациями, накапливаемых и развиваемых сообществами исследователей. Средства проведения экспериментов в инфраструктурах исследований также поддерживаются концептуальными спецификациями, которые обеспечивают основу для проведения измерений, изучения свойств сущностей предметной области, применения методов данной предметной области, описания и проверки гипотез. На примере предметной области астрономии показаны принципы построения концептуальных спецификаций и их использования при анализе данных.

Работа выполнена при частичной поддержке РФФИ (гранты 16-07-01028, 16-07-01162, 15-29-06045, 14-07-00548).

Труды XVIII Международной конференции DAMDID/RCDL'2016 «Аналитика и управление данными в областях с интенсивным использованием данных», Ершово, 11-14 октября 2016

1 Введение

Исследовательские задачи критически зависят от растущих и дополняющих одна другую массивных коллекций данных, собираемых в результате наблюдений, экспериментов и моделирования. Одновременно растёт качество данных и соответственно глубина требуемого анализа данных. Подходы к исследованиям, при которых для решения задач производился выбор источников данных и формулирование задач в их терминах, стали трудоёмкими при множестве неоднородных источников данных и большом количестве способов их анализа. Если программы, реализующие решение задач анализа данных, зависят от конкретных источников данных, это препятствует масштабированию для неоднородных и массивных источников данных, накоплению реализаций методов анализа данных, их интероперабельности и повторному использованию в различных исследованиях [1].

От поиска и связывания источников данных для решения поставленных задач акцент исследований смещается в направлении анализа доступных массивных коллекций данных для нахождения новых знаний в предметной области исследования [2]. Разрабатываются научные методы оценки характеристик объектов по наблюдаемым параметрам, методы обобщения, классификации, выявления и исследования интересующих сущностей и явлений, средства генерации и проверки научных гипотез, специализированные процедуры в определённых направлениях науки, обеспечивается их автоматизированное применение над данными массивных коллекций и доступность для сообществ, работающих в инфраструктурах исследований.

Для разностороннего изучения конкретных типов сущностей реального мира оказывается важным совместное использование средств исследований и концептуальных спецификаций, определяющих как семантику сущностей и явлений в предметной области, так и семантику применяемых в ней

методов. Поэтому одной из задач сообществ, занимающихся исследованиями в определённой предметной области, является концептуализация предметной области для построения таких спецификаций и связывания с ними данных, реализаций методов и процессов.

Для обеспечения науки методами и средствами, применимыми к объектам предметных областей, в работе предлагается подход к концептуализации предметных областей для их исследования. Он опирается на явное описание семантики сущностей и процессов при формулировании постановок и реализаций алгоритмов решения научных задач, обеспечивая их семантическое соответствие спецификациям предметной области. В свою очередь, различные источники данных, в том числе научных данных, семантически отображаются в концептуальные спецификации предметной области исследования. Аналитические задачи формулируются также в терминах концептуальных спецификаций предметной области и решаются с использованием отображённых в них данных и методов.

В настоящей работе показано, каким образом концептуальные спецификации предметных областей, поддерживаемые заинтересованными сообществами, могут быть использованы для организации исследований. В статье для этого используются концептуальные спецификации предметных областей из области астрономии. Следующий раздел посвящён принципам определения спецификаций предметных областей исследований. В разделе 3 описаны подходы к накоплению научных методов и перспективы построения инфраструктур исследований на основе коллекций методов. Раздел 4 посвящён использованию спецификаций предметных областей для проведения экспериментов, описания и проверки научных гипотез, организации потоков работ в инфраструктурах исследовательских сообществ для манипулирования данными и методами при проведении экспериментов.

2 Средства спецификации предметных областей

Процесс концептуализации предметных областей, в первую очередь, предполагает разработку онтологий в исследовательских сообществах для формализации и систематизации знаний о характерных для этих областей сущностях и явлениях. Члены сообществ действуют в рамках онтологического обязательства, определённого такими онтологиями, то есть используют понятия предметных областей непротиворечивым образом по отношению к теориям, специфицируемым онтологиями. Для обеспечения такого подхода важна автоматизация контроля непротиворечивости результатов действий при любых манипуляциях понятиями предметной области.

Концептуальные определения предметных областей для проведения исследований включают следующие описания:

- понятия сущностей, фигурирующих в предметной области в качестве исследуемых или связанных с ними;
- понятия, определяющие характеристики и поведение объектов предметной области;
- понятия, соответствующие научным методам, корреляциям, существующим в предметной области исследования;
- понятия, определяющие подходы к наблюдению объектов и моделированию сущностей предметной области, проведению научных экспериментов.

Языковые средства представления формальных онтологий включают понятия, отношения понятий и ограничения, связанные с понятиями, обычно выраженные в подмножестве логики предикатов, отображаемом в некоторую дескриптивную логику или другие формальные модели. Скалярные типы данных в онтологиях предпочтительно не использовать, так как они отражают некоторые отношения, которые на онтологическом уровне лучше описывать явно для однозначной интерпретации понятий. Также в онтологиях традиционно не используются средства спецификации методов. Однако ограничения, связанные с поведением объектов предметной области, необходимо специфицировать и в онтологии. Это делается посредством определения понятий, соответствующих разного рода корреляциям характеристик сущностей и процессам.

Процесс концептуализации предметной области помимо определения понятий включает разработку концептуальных схем предметных областей, отличающихся от онтологий, в первую очередь, своим назначением [24]. Они определяют не просто понятия предметной области, а структуры представления информации об объектах предметных областей и спецификации поведения для манипулирования объектами. Однако, если разработаны онтологии, то концептуальные схемы составляются согласно знаниям об сущностях, зафиксированным в этих онтологиях. Принципы составления концептуальных схем предметных областей на основе определений онтологий описаны в [3]. Языковые средства спецификации концептуальных схем включают определения абстрактных типов данных, представляющие информацию о состоянии объектов и характеризующихся наборами атрибутов, значения которых соответствуют определённым типам данных от простых скалярных до объектных типов и ассоциаций. С типами и атрибутами типов могут быть связаны метаданные, определяющие их собственные характеристики. Множества однотипных объектов могут составлять классы. Поведение объектов предметной области выражается методами типов.

Любые структуры или информационные объекты целесообразно сопровождать метаинформацией о том, с какими понятиями онтологии они связаны, чтобы фиксировать их семантику и систематизировать в соответствии с ней ресурсы, имеющиеся в арсенале исследователей определённой предметной области.

Формальность спецификации онтологий и концептуальных схем принципиально важна для обеспечения семантической интеграции информационных ресурсов и воспроизводимости программ над спецификациями предметной области. Без доказательного подхода использование спецификаций предметной области мало отличается от умозрительного связывания элементов схем при интеграции ресурсов. Формального обоснования в концептуальном подходе требуют такие задачи, как, например:

- проверка внутренней непротиворечивости спецификаций онтологий и концептуальных схем предметной области:
- контроль интероперабельности совмещаемых или замещаемых спецификаций;
- проверка соответствия разрабатываемых спецификаций концептуальных схем знаниям, отражённым в онтологии;
- обнаружение спецификаций информационных ресурсов, семантически соответствующих спецификациям предметной области;
- проверка соответствия используемых информационных ресурсов спецификациям предметной области.

От выбранного формализма языка спецификации зависит возможность применения средств автоматического вывода. В частности, дескриптивные логики используются в качестве основ диалектов языка онтологий OWL. Для спецификаций, приводимых к дескриптивным логикам, перечисленные выше задачи разрешимы.

Спецификации в моделях, основанных на логиках, целесообразно представлять в унифицированном виде, в частности, в диалектах языка RIF [17]. Помимо прочего, язык RIF может использоваться для выражения правил над спецификациями на языке OWL, что позволяет определять формальные спецификации поведения объектов предметной области и алгоритмы решения задач напрямую над онтологиями OWL. В мультидиалектной архитектуре в зависимости от используемых диалектов RIF для рассуждений над спецификациями используются соответствующие им системы вывода [23].

Для языков, выразимых в логике предикатов первого порядка, те же задачи могут быть решены в интерактивном режиме при помощи доказательства уточнения спецификаций программ [18]. Уточняющая спецификация может быть использована вместо уточняемой. В частности, при разработке спецификаций концептуальных схем предметных областей на основе онтологии необходимо, чтобы онтология уточнялась спецификациями схем. Для обоснования этого язык онтологий и язык концептуальных схем должны быть отображены в язык абстрактных машин системы вывода B, обеспечивающей доказательство уточнения [19-21]. При этом понятия, определяющие зависимости и процессы, отображаются в операции, которые должны уточняться операциями типов в концептуальной схеме. Данные о том, какие элементы схемы сформированы в соответствии с понятиями онтологии, сохраняются в схеме в качестве метаданных. Таким образом, явно специфицируется семантика элементов схем с точки зрения понятий предметной области.

Онтологии и концептуальные схемы предметных областей разрабатываются и поддерживаются сообществами, работающими в этих областях, таким образом, чтобы быть достаточными для нужд научных групп. Средства и состав концептуальных спецификаций предметных областей в сообществе направлены на семантическую интероперабельность взаимодействующих компонентов, повторное использование информационных ресурсов и воспроизводимость программ за счёт привязки к семантике предметной области. Поэтому как на уровне онтологий, так и на уровне концептуальных схем предъявляются высокие требования к полноте и формальности спецификаций.

Оценивая использование концептуальных спецификаций на примере области астрономии, необходимо отметить, что в рамках альянса Международной виртуальной обсерватории разрабатываются соответствующие стандарты. Известны онтологии [4, 5], однако они созданы на основе тезаурусов и не содержат многих существенных понятий и отношений, которые необходимы для работы исследователей, не отражают ограничений состояния и поведения объектов, явлений и научных экспериментов в предметной области. Нет хорошо формализованных онтологий, направленных на логический вывод.

К концептуальным спецификациям в астрономии относятся также разрабатываемые стандарты концептуальных схем наиболее общих областей, которые затрагиваются практически в каждой астрономической задаче, в частности:

- Space-Time Coordinate Metadata [6] – схема свойств различных систем координат;
- Photometry Data Model [7] – схема и формат сериализации фотометрической информации, определяющий также функции калибровки и преобразования между разными фотометрическими системами;
- VOEvent [8] – схема описания наблюдаемых объектов и астрофизических явлений, включающая идентификацию объектов, наблюдателей, место, время и средства наблюдения.

Схемы, как и онтологии Международной виртуальной обсерватории, не описывают объекты и научные методы специфических областей и, в основном, не включают ограничений целостности и спецификаций поведения объектов.

Концептуализация необходима не только в наиболее общих областях, затрагиваемых астрономией, таких как фотометрия и спектроскопия, но и в областях, представляющих интерес для более узких кругов исследователей, а также на границе между областями, где чаще всего возникает сотрудничество научных коллективов и повторное использование результатов исследований. При этом важно описание как объектов исследования, так и методов исследования и проведения экспериментов в таких областях. Прототип[1] разработанной авторами статьи онтологии в области астрономии, определяющей некоторые специфические области наряду с общеупотребимыми понятиями, представлен на языке OWL [9].

Модульная структура онтологии, по сути, описывает взаимодействующие предметные области в рамках астрономии. Она включает области, определяемые разными объектами исследования, методами наблюдения и моделирования, подходами к исследовательскому процессу в целом. По мере расширения затрагиваемых областей и круга задач, решаемых взаимодействующими группами исследователей, в онтологии развиваются различные модули.

Разработана онтология для спецификации научных экспериментов, формирующая междисциплинарные базисные понятия. Она включает: 1) онтологию характеристик измерений объектов исследования, включающую такие понятия как единицы измерений, погрешности, законы распределения значений и другие; 2) онтологию взаимозависимостей измерений объектов, необходимую для введения понятийного аппарата для спецификации поведения объектов и включающую понятие корреляции измерений и его подпонятия, определяющие понятия функции, метода, закона, гипотезы и другие.

Рассмотрим некоторые спецификации на языке OWL. В части онтологии, определяющей понятия, используемые для проведения научных экспериментов, определены несколько модулей. Среди них модуль, содержащий базовые понятия, относящиеся к измерениям параметров исследуемых объектов, включает понятие измерения (Measurement), связываемое с объектом исследования (AstrObject) отношением isMeasurementOf, понятия значений параметров, единиц измерений, точности измерений, характеризуемой статистической и систематической ошибками.

```
Class(Measurement
  restriction(hasValue
    allValuesFrom(Value))
  restriction(hasUnit
    allValuesFrom(MeasurementUnit))
  restriction(hasError
    allValuesFrom(MeasurementError)
    maxCardinality(1))
  restriction(isMeasurementOf
    allValuesFrom(AstrObject)
    maxCardinality(1)))
Class(MeasurementUnit
  restriction(hasScaleFactor
    allValuesFrom(ScaleFactor)
    maxCardinality(1))
  restriction(hasProjection
    allValuesFrom(ScaleProjection)
    maxCardinality(1)))
Class(MeasurementError
  restriction(isErrorOf
    allValuesFrom(Measurement)
    maxCardinality(1)))
Class(StatisticalError
  partial MeasurementError)
Class(SystematicError
  partial MeasurementError)
Class(Value
  restriction(isValueOf
    allValuesFrom(Measurement)
    maxCardinality(1)))
ObjectProperty(isValueOf
  domain(Value)
  range(Measurement)
  inverseOf(hasValue))
ObjectProperty(isMeasurementOf
  domain(Measurement)
  range(AstrObject)
  inverseOf(hasMeasurement))
ObjectProperty(isErrorOf
  domain(MeasurementError)
  range(Measurement)
  inverseOf(hasError))
```

Модуль астрономических объектов определяет понятия, связанные с характеристиками, общими для произвольных астрономических объектов. Спецификация понятия астрономического объекта (AstrObject) включает связи с другими понятиями онтологии: его координатами, измерениями разного рода физических параметров, связью с составными объектами, к которым данный объект принадлежит в качестве компонента, и другими:

```
Class(AstrObject
  restriction(hasIdentifier
    allValuesFrom(Identifier))
  restriction(hasCoordinate
    allValuesFrom(Coordinate))
  restriction(inEpoch
    allValuesFrom(Epoch)
    maxCardinality(1))
  restriction(hasMeasurement
    allValuesFrom(Measurement))
  restriction(hasMorphology
    allValuesFrom(Morphology)
    maxCardinality(1))
  restriction(hasProcess
    allValuesFrom(Process))
  restriction(isComponentOf
    allValuesFrom(CompoundObject)))
```

Онтологический модуль, описывающий предметную область звёзд, включает понятие звёздного объекта (StellarObject) как точечной сущности в Галактике, самостоятельного или являющегося компонентом составного объекта,

[1] http://ontology.ipi.ac.ru/ontologies/astront/

понятие звезды (Star) как одиночного звёздного объекта, понятие кратной звезды как звёздного объекта, состоящего из компонентов, а также ряд специфических понятий характеристик звёзд:

```
Class(StellarObject
  partial AstrObject
  restriction(hasMorphology
    hasValue(PointObject)))
Class(Star
  partial StellarObject)
```

Конкретные виды измерений определяются как подпонятия понятия Measurement в специализированных модулях онтологии, в которых они используются. Модуль астрофизических параметров астрономических объектов содержит общие физические характеристики объектов, такие как температура, масса, размеры, светимость. В частности, масса является общей характеристикой астрономических объектов:

```
Class(Mass
  partial Measurement)
```

Представим себе понятие массы звезды (StarMass), являющееся подпонятием понятия масса (Mass). Оно использует понятия разных модулей, ограничивая тип описываемых сущностей как звёзды и определяя в качестве единицы измерения массу Солнца.

```
Class(StarMass
  partial Mass
  restriction(isParameterOf
    allValuesFrom(Star))
  restriction(hasUnit
    hasValue(SunMass)))
```

При переходе от онтологии к концептуальной схеме необходимо сформировать структуры для представления информации об объектах предметной области. В мультидиалектной архитектуре помимо спецификаций OWL используется язык СИНТЕЗ [10] для реализации на основе применения предметных посредников. В то время как OWL является разрешимым языком для задачи включения, для языка СИНТЕЗ отработано доказательство уточнения спецификаций. В представление на языке СИНТЕЗ также могут быть отображены спецификации в диалекте RIF BLD. Также разработано отображение языка OWL в язык СИНТЕЗ [11].

Пример спецификации концептуальной схемы на языке СИНТЕЗ, построенной в соответствии с онтологической спецификацией, определяет структуру представления информации о звёздах с атрибутом, хранящим массу звезды и метаданными, определяющими единицу её измерения в массах Солнца:

```
{ Star;
  in: type;
  crd: Coordinate;
  mass: Float;
    metaslot
      in: measurement;
      hasUnit: SunMass;
    end
}
```

Спецификации концептуальных схем требуют также определения методов и функций. Такие спецификации формируются в схеме на основе понятий онтологии, описывающих зависимости характеристик объектов, а также понятий процессов. Они рассматриваются в следующем разделе.

3 Организация коллекций научных данных и методов

Сегодня активно развиваются библиотеки методов в специализированных областях исследований в астрономии, в инфраструктурах совместных исследований, где помимо данных в доступ исследовательскому сообществу предоставляются всевозможные сервисы, а также средства их поиска и описания для правильного использования.

Одной из первых систем, предоставляющих технологии для работы сообществ в области астрономии была сеть AstroGrid [12]. Она представляла собой инфраструктуру для решения задач виртуальной обсерватории и состояла из множества узлов, содержащих всевозможные сервисы и ресурсы. Архитектура AstroGrid включала реестр, представляющий собой коллекцию метаданных, описывающих ресурсы, которые могут использоваться при решении задач. Это позволяло организовать поиск доступных коллекций данных и методов по метаданным. Проект был закрыт, в первую очередь, по причине медленного развития сети. Организация узлов сети оказалась сложной для широкого распространения в астрономическом сообществе.

Проект WF4Ever [13] направлен на сохранение результатов научных исследований над данными, с этой целью разработаны средства курирования объектов исследования как комплексных объектов, включающих документы, данные, сервисы, потоки работ. В рамках этого же проекта развивается библиотека сервисов [14], которые обеспечивают доступ к существующим астрономическим веб-сервисам и к данным каталогов, преобразование между разными стандартными представлениями и манипулирование таблицами при соединении разных источников данных, и используются в качестве элементов потоков работ. Библиотека получила признание научного сообщества, в первую очередь, за счёт простоты построения процессов обработки данных без программирования.

В проекте EUDAT [22] ставится задача построения инфраструктуры доступа к научным данным. Семантические подходы к её организации включают ведение репозиториев словарей, включающих термины широкого круга научных областей. Помимо словарей общего назначения, определяющих такие атрибуты как название, авторы, научная дисциплина, определяются иерархии терминов, именующих научные дисциплины, научные методы и объекты. Специфические для предметных областей словари разрабатываются научными сообществами. Эти описания используются для организации информационно-поисковой системы, обеспечивающей поиск

релевантных задаче сервисов. Поиск может производиться одновременно по терминам разных словарей на пересечении исследований разных научных сообществ. Европейская виртуальная обсерватория представлена в проекте как одна из областей исследования, и её решения сравниваются с подходами EUDAT.

В ближайшей перспективе в области астрономии появятся источники данных, объёмы которых намного превышают сегодняшние. Такие проекты как широкоугольный телескоп LSST и космическая обсерватория Gaia будут генерировать потоковые данные наблюдений. Для их обработки заранее готовятся каналы передачи данных для их локализации в местах исследований, решаются вопросы доступа к данным различных научных учреждений, а также разрабатываются общедоступные средства обработки данных и средства их эффективного поиска [15].

Для эффективного взаимодействия внутри сообщества, имеющего доступ к данным, и во избежание появления множество разрозненных работ кооперация исследователей в подобных проектах должна основываться на обеспечении доступа к разработке планов исследований, специализированным методам и результатам анализа данных. Таким образом, помимо накопления данных, необходимо накопление доступных методов, алгоритмов и инструментов обработки, готовых к применению над большими массивами данных. Концептуализация предметной области в рамках сообщества и семантические подходы позволяют систематизировать методы предметной области в исследовательских инфраструктурах. И научные данные, и научные методы связываются со спецификациями предметных областей, к которым они относятся.

Во всякой предметной области накапливаются знания и законы предметной области, специфические методы, направленные на определённые виды анализа сущностей, фигурирующих в предметной области. Помимо этого, должен быть доступен широкий круг аналитических методов и инструментов общего назначения, применяемых вне рамок специфической предметной области. К таким методам относятся, например, численные, статистические методы, методы машинного обучения и другие.

В рамках онтологических моделей, традиционно не имеющих средств спецификации методов, концептуализация поведения объектов может определяться понятиями, отражающими зависимые характеристики объектов и процессы. Одним из модулей разрабатываемой онтологии является модуль, определяющий взаимозависимости измерений. Под понятием корреляции (Correlation) подразумевается корреляция определённых параметров измерения у объектов предметной области:

```
Class(Correlation
  restriction(isCorrelationOf
    allValuesFrom(Measurement))
  restriction(hasRegression
    allValuesFrom(RegressionFunction))
  restriction(hasRMSDeviation
    allValuesFrom(RMSDeviation))
  restriction(isCausal
    allValuesFrom(TruthValue)))
Class(Hypothesis
  partial Correlation
  restriction(explains
    allValuesFrom(Phenomenon))
  restriction(derivedFrom
    allValuesFrom(Hypothesis))
  restriction(competesWith
    allValuesFrom(Hypothesis))
  restriction(hasProbability
    allValuesFrom(Probability))
  restriction(hasPValue
    allValuesFrom(Probability))
  restriction(hasQuality
    allValuesFrom(TruthValue)))
Class(Law
  partial Hypothesis
  Restriction(hasQuality
    hasValue(True)))
```

Посредством понятий на уровне онтологии декларативно описываются научные методы, гипотезы, законы, модели, процессы, эксперименты, связанные с характеристиками объектов предметной области. Понятие гипотезы определяется как разновидность статистически подтверждаемых корреляций. Для статистического подтверждения гипотез, с одной стороны, используется моделирование, обеспечивающее их математическое описание, а с другой стороны, – эксперимент для сравнения результатов моделирования с данными наблюдения объектов исследуемых объектов.

Понятия научных методов, законов и гипотез определяются как подпонятия корреляции измерений с указанием ограничений конкретных зависимых величин. Их понятийное описание не зависит от конкретных реализаций и представлений, будь то таблиц значений или коэффициентов, точных математических формул, функций распределения, программ или других возможных способов описания.

Рассмотрим спецификации концептуализации гипотезы начальной функции масс в составе Безансонской модели Галактики [25] на основе онтологических спецификаций предметной области, приведённых выше. Эта гипотеза связана с предположением о достаточно постоянном распределении звёзд разной массы в некотором ограниченном объёме пространства Галактики. Другими словами, гипотеза предполагает зависимость количества звёзд от их массы в фиксированном объёме пространства, которому принадлежат эти звёзды:

```
Class(InitialMassFunction
  partial Hypothesis
  restriction(isCorrelationOf
    ObjectSomeValuesFrom(StarMass))
  restriction(isCorrelationOf
    ObjectSomeValuesFrom(
      intersectionOf(
        Quantity
        restriction(hasElement
          allValuesFrom(Star))))))
```

На основе онтологии коллекции различных реализаций методов могут быть систематизированы по различным признакам, соответствующим понятиям онтологий: исследуемым объектам, характеристикам объектов, свойствам, зависимым от данной характеристики, известным методам и гипотезам и другим. Соответственно, по любому из таких признаков исследователями может производиться поиск существующих реализаций научных методов для их повторного использования.

4 Средства проведения научных экспериментов

Применение методов, собранных в коллекции, при исследованиях в предметной области происходит в соответствии с определёнными сценариями. Так, автоматизированный запуск анализа данных может происходить при появлении данных с определёнными свойствами или определённого типа для обогащения данных об объектах определёнными характеристиками, которые в свою очередь могут использоваться для дальнейших исследований.

Методика исследования обычно состоит из определённых шагов, включающих очистку и анализ данных, построение научных гипотез, моделирование в соответствии с гипотезами и проверку моделей на данных наблюдений. Эксперименты над данными формулируются на основе создания новых методов и повторного использования существующих реализаций методов в спецификациях потоков работ.

Инфраструктуры поддержки научных исследований, помимо возможности использования коллекций данных и реализаций методов в определённых предметных областях, должны содержать средства проведения научных экспериментов. В частности, это касается возможности формулирования и проверки научных гипотез.

Использование концептуальных спецификаций при формулировании и тестировании гипотез даёт те же преимущества, что и при управлении коллекциями методов и решении научных задач над ними.

На уровне концептуальных спецификаций понятиям методов и законов и гипотез приводятся в соответствие методы и правила. Спецификации ограничений понятий зависимых величин уточняются предусловиями и постусловиями методов.

Так на основе знаний, специфицированных в понятии, описывающем гипотезу начальной функции масс могут быть созданы абстрактные типы данных концептуальной схемы для моделирования и проверки гипотез. Тип будет включать определение интерфейса функции с параметрами, соответствующими отношениям в понятии зависимости.

```
{ IMF;
  in: type;
  supertype: Hypothesis;
  draw_mass: { in: function;
    params: { +mass/Real, –quantity/Real } };
}
```

Если в онтологической спецификации определялось направление зависимости (функция), в соответствии с ним определяются и входные и выходные параметры. Иначе выбор направления определяется нуждами решаемой задачи. В предусловии и постусловии функции определяются ограничения, соответствующие онтологическим определениям. Ограничения в онтологии, связанные с измерениями, единицами измерений, точностью измерений, помогают сформировать структуру метаданных для сопровождения измерений в концептуальной схеме. Таким образом, концептуальные спецификации предметной области используются для моделирования закономерностей Галактики и исследования моделей.

Представленный абстрактный тип данных может быть реализован разными способами для организации проверки гипотезы. Существующие реализации моделей, соответствующих гипотезе, могут быть найдены в коллекции методов по онтологическим описаниям. Для конкретной реализации определяется подтип данной спецификации.

С другой стороны, та же спецификация типа используется для проверки гипотезы на данных экспериментов. Для этого используются данные из множества источников, которые интегрированы в концептуальную схему предметной области. Подтип вышеприведённой спецификации строится таким образом, чтобы по входным данным получать данные наблюдения, соответствующие выходным параметрам. Проверка гипотезы производится с помощью сравнения результатов моделирования с результатами, полученными на данных реальных источников.

Реализация моделей и экспериментов может использовать доступные методы общего назначения, такие как методы машинного обучения или численные методы, однако спецификации онтологий и схем, принятых в сообществе, от них не зависят.

Эффективность исследований, проводимых сообществом предметной области, зависит не только от доступности данных наблюдения, реализаций методов и моделей, но также от планирования экспериментов, в котором учитываются онтологические знания об изучаемых объектах и взаимозависимостях их характеристик (гипотезах и законах).

Концептуальные спецификации могут быть использоваться при генерации гипотез. Генерация на основе поиска корреляций в данных требует проверки семантического соответствия коррелирующих параметров объектов. Не связанные друг с другом хотя бы опосредованно параметры в спецификациях понятий с меньшей долей вероятности рассматриваются как коррелирующие.

При взаимодействии разных гипотез в одной модели взаимное влияние их параметров, участвующих в гипотезах, может быть учтено на основе знаний онтологии о зависимости друг от друга разных измерений. Эти зависимости могут быть исследованы и их реализации найдены посредством семантического поиска с использованием онтологии. Ограничения концептуальных схем при этом могут гарантировать согласованность модели.

Для моделирования и проверки гипотез над концептуальными схемами разрабатываются потоки работ, которые реализуют процесс моделирования, проверки гипотез, их корректировки для подбора параметров моделей, наилучшим образом повторяющих результаты, полученные на реальных данных.

Заключение

В статье рассмотрены вопросы концептуализации предметных областей для организации научных исследований над данными. Развитие инфраструктур поддержки научных исследований, в основе которых лежат концептуальные спецификации предметных областей, развиваемые и поддерживаемые сообществами, работающими в этих областях, позволяет избежать зависимости программ от структуры источников данных, обеспечить интероперабельность различных методов при совместной работе, повысить надёжность результатов за счёт использования формальных непротиворечивых спецификаций. Рассмотрены возможности концептуального анализа предметной области для формализации научных гипотез и их тестирования на основе данных наблюдений.

Литература

[1] Д. О. Брюхов, А. Е. Вовченко, Л. А. Калиниченко. Поддержка повторного использования спецификаций потоков работ за счет обеспечения их независимости от конкретных коллекций данных и сервисов // Всероссийская конференция «Электронные библиотеки» RCDL 2013. – CEUR Workshop Proceedings, 2013. – Т. 1108. – С. 61-69.

[2] The Fourth Paradigm: Data-Intensive Scientific Discovery. T. Hey, et al (Eds). – Microsoft Research. – Redmond, 2009.

[3] А. Е. Вовченко, В. Н. Захаров, Л. А. Калиниченко и др. От спецификаций требований к концептуальной схеме // Труды 12-й Всероссийской научной конференции Электронные библиотеки: перспективные методы и технологии, электронные коллекции RCDL 2010. – Казань: КФУ, 2010. – С. 375-381.

[4] Ontology of Astronomical Object Types. Version 1.20. – IVOA, 2009. – URL: http://www.ivoa.net/documents/latest /AstrObjectOntology.html

[5] IVOAO Ontology. – University of Maryland, 2010. – URL: http://www.astro.umd.edu/~eshaya/astro-onto/

[6] Space-Time Coordinate Metadata for the Virtual Observatory
Version 1.33. – IVOA, 2011. – URL: http://www.ivoa.net/documents/ latest/STC.html

[7] IVOA Photometry Data Model. Version 1.0. – IVOA, 2013. – URL: http://www.ivoa.net/documents/PHOTDM/

[8] Sky Event Reporting Metadata (VOEvent). Version 2.0. – IVOA, 2011. – URL: http://www.ivoa.net/Documents/VOEvent/

[9] OWL 2 Web Ontology Language. Document Overview (Second Edition).} -- W3C, 2012. -- URL: http://www.w3.org/TR/owl-overview/

[10] L. A. Kalinichenko, S. A. Stupnikov, D. O. Martynov. SYNTHESIS: a Language for Canonical Information Modeling and Mediator Definition for Problem Solving in Heterogeneous Information Resource Environments. Moscow: IPI RAN, 2007. – 171 p.

[11] L. A. Kalinichenko, S. A. Stupnikov. OWL as Yet Another Data Model to be Integrated. Advances in Databases and Information Systems: Proc. II of the 15th East-European Conference. – Vienna: Austrian Computer Society, 2011. – P. 178-189.

[12] AstroGrid. – URL: http://www.astrogrid.org/

[13] K. Belhajjame, et al. Workflow-Centric Research Objects: A First Class Citizen in the Scholarly Discourse // ESWC2012 Workshop on the Future of Scholarly Communication in the Semantic Web (SePublica2012). – Heraklion, 2012.

[14] N. A. Walton, et al. Taverna and workflows in the virtual observatory // Astronomical Data Analysis Software and Systems ASP Conference Series. – Vol. 394. – 2007. – P. 309.

[15] M. Luric, T. Tysoc. LSST Data Management: Entering the Era of Petascale Optical // Astronomy. Highlights of Astronomy. – Vol. 16. – 2015. – P. 675.

[16] N. A. Skvortsov, et al. Conceptual approach to astronomical problems // Astrophysical Bulletin. – Vol. 71, No. 1. – Springer, 2016.

[17] Rule interchange format: The framework // Web Reasoning and Rule Systems: 2nd Conference (International) Proceedings, LNCS 5348. – Berlin–Heidelberg: Springer Verlag, 2008. – P. 1-11.

[18] Abrial J.R. The B Book - Assigning Programs to Meanings. - Cambridge: Cambridge University Press, 1996.

[19] Н. А. Скворцов. Применение уточнения понятий в решении задач манипулирования онтологиями // RCDL'2007. – Переславль-Залесский: УГП, 2007. – С.225-229.

[20] Н. А. Скворцов. Использование системы интерактивного доказательства для

отображения онтологий // RCDL'2006. – Ярославль: ЯрГУ, 2006. – С. 65-69.

[21] С. А. Ступников. Отображение спецификаций, выраженных средствами ядра канонической модели, в язык AMN // Системы и средства информатики: Спец. вып. Формальные методы и модели в композиционных инфраструктурах распределенных информационных систем. Под ред. И. А. Соколова. М.: ИПИ РАН, 2005. С. 69-95.

[22] H. Schentz, Y. le Franc. Building a semantic repository using B2SHARE // EUDAT 3rd Conference. – 2014.

[23] L. A. Kalinichenko, S. A. Stupnikov, E. A. Vovchenko, D. Y. Kovalev. Rule-based Multi-dialect Infrastructure for Conceptual Problem Solving over Heterogeneous Distributed Information Resources // Advances in Intelligent Systems and Computing. – Springer, 2013. – V. 241. – P. 61-68.

[24] М. Р. Когаловский, Л. А. Калиниченко. Концептуальное моделирование в технологиях баз данных и онтологические модели. // Онтологическое моделирование: состояние и направления исследований и применения. - М. ИПИ РАН, 2008.

[25] A. C. Robin, C. Reylé, S. Derrière and S. Picaud. A synthetic view on structure and evolution of the Milky Way, 2003, Astron. Astrophys., 409:523 ADS

Conceptual modeling of subject domains in data intensive research

Nikolay A. Skvortsov, Leonid A. Kalinichenko, Dmitry Yu. Kovalev

Nowadays research of various scopes especially in natural sciences requires manipulation of big volumes of data generated by observation, experiments and modeling. Organization of data-intensive research assumes definition of domain specifications including concepts (specified by ontologies) and formal representation of data describing domain objects and their behavior (using conceptual schemes), shared and maintained by communities working in the respective domains. Research infrastructures are based on domain specifications and provide methods applied to such specifications, collected and developed by research communities. Tools for organizing experiments in research infrastructures are also supported by conceptual specifications of measuring and investigating object properties, applying of research methods, describing and testing of hypotheses. Astronomy as a sample data intensive domain (DID) is chosen to demonstrate building of conceptual specifications and usage of them for data analysis.

Rule-based inference of data lineage, impact and semantic models

© Kalle Tomingas © Priit Järv © Tanel Tammet
Tallinn University of Technology, Ehitajate tee 5, Tallinn 19086
kalle.tomingas@gmail.com priit.jarv@gmail.com tanel.tammet@gmail.com

Abstract

The paper presents methods to calculate meaningful data transformation and component dependency paths purely from the database structure and associated procedures and queries, using heuristic impact analysis, probabilistic rules and semantic technologies. The dependencies are categorized, aggregated and visualized to address various planning and decision support problems.

1 Introduction

Developers and managers are facing similar data lineage and impact analysis problems in complex data integration (DI), business intelligence and Data Warehouse (DW) environments where the chains of data transformations are long and the complexity of structural changes is high. The management of data integration processes becomes unpredictable and the costs of changes can be very high due to the lack of information about data flows and internal relations of system components. Important contextual relations are coded into data transformation queries and programs (e.g. SQL queries, data loading scripts, open or closed DI system components etc.). Data lineage dependencies are spread between different systems and frequently exist only in program code or SQL queries. This leads to unmanageable complexity, lack of knowledge and a large amount of technical work with uncomfortable consequences like unpredictable results, wrong estimations, rigid administrative and development processes, high cost, lack of flexibility and lack of trust.

We point out some of the most important and common questions for large DW environments which usually become a topic of research for system analysts and administrators:

- Where does the data come or go to in/from a specific column, table, view or report?
- When was the data loaded, updated or calculated in a specific column, table, view or report?
- Which components (reports, queries, loadings and structures) are impacted when other components are changed?
- Which data, structure or report is used by whom and when?
- What is the cost of making changes?
- What will break when we change something?

The ability to find ad-hoc answers to many day-to-day questions determines not only the management capabilities and the cost of the system, but also the price and flexibility of making changes.

To address these research questions, we proposed a data lineage and impact analysis system [15]. We will build on that foundation and offer a further refinement of the inference of information based on the data collected by analysis of the SQL statements.

2 Related Work

Impact analysis, traceability and data lineage issues are not new. A good overview of the research activities of the last decade is presented in an article by [11]. We can find various research approaches and published papers from the early 1990's with methodologies for software traceability [12]. The problem of data lineage tracing in data warehousing environments has been formally founded by Cui and Widom [3,4]. Our recent papers build background to the theory by introducing the Abstract Mapping representation of data transformations and rule-based impact analysis [16].

Other theoretical works for data lineage tracing can be found in [6] and [7]. Fan and Poulovassilis developed algorithms for deriving affected data items along the transformation pathway [6]. These approaches formalize a way to trace tuples (resp. attribute values) through rather complex transformations, given that the transformations are known on a schema level. This assumption does not often hold in practice. Transformations may be documented in source-to-target matrices (specification lineage) and implemented in ETL tools (implementation lineage). Woodruff and Stonebraker create solid base for the data-level and the operators processing based the fine-grained lineage in contrast to the metadata based lineage calculation in their research paper [19].

Other practical works that are based on conceptual models, ontologies and graphs for data quality and data lineage tracking can be found in [14], [17] and [18]. De Santana proposes the integrated metadata and the CWM metamodel based data lineage documentation approach [5]. Priebe et al. [11] concentrates on proper handling of specification lineage, a huge problem in large-scale DWH projects, especially in case different sources have

Proceedings of the XVIII International Conference «Data Analytics and Management in Data Intensive Domains» (DAMDID/RCDL'2016), Ershovo, Russia, October 11 - 14, 2016

to be consistently mapped to the same target. They propose a business information model (or conceptual business glossary) as the solution and a central mapping point to overcome those issues.

Scientific workflow provenance tracking is closely related to data lineage in databases. The distinction is made between coarse-grained, or schema-level, provenance tracking [8] and fine-grained or data instance level tracking [10]. The methods of extracting the lineage are divided to physical (annotation of data, such as in [10]) and logical, where the lineage is derived from the graph of data transformations [9].

In the context of our work, efficiently querying of the lineage information after the provenance graph has been captured, is of specific interest. Heinis and Alonso [8] present an encoding method that allows space-efficient storage of transitive closure graphs and enables fast lineage queries over that data. Anand et al. [2] propose a high level language QLP, together with the evaluation techniques that allow storing provenance graphs in a relational database.

3 System Architecture

The inference method of the data flow and the impact dependencies that presented in this paper is part of a larger framework of a full impact analysis solution. The core functions of the system architecture are built upon the following components presented in the Figure 1 and described in detail in [15],[16].

Figure 1 Impact Analysis system architecture components.

The core functions of the system architecture are built upon the following components in the Figure 1:

1. Scanners collect metadata from different systems that are part of DW data flows (DI/ETL processes, data structures, queries, reports etc.).
2. The SQL parser is based on customized grammars, GoldParser[1] parsing engine and the Java-based XDTL engine.
3. The rule-based parse tree mapper extracts and collects meaningful expressions from the parsed text, using declared combinations of grammar rules and parsed text tokens.
4. The query resolver applies additional rules to expand and resolve all the variables, aliases, sub-query expressions and other SQL syntax structures which encode crucial information for data flow construction.
5. The expression weight calculator applies rules to calculate the meaning of data transformation, join and filter expressions for impact analysis and data flow construction.
6. The probabilistic rule-based reasoning engine propagates and aggregates weighted dependencies.
7. The open-schema relational database using PostgreSQL for storing and sharing scanned, calculated and derived metadata.
8. The directed and weighted sub-graph calculations, and visualization web based UI for data lineage and impact analysis applications.

3.1 Weight Estimation

In the stages preceding the inference engine, data structure transformations are parsed, extracted from queries and stored as formalized, declarative mappings in the system. The SQL parsing stages store the following metadata for each mapping:

- *StrList* - String constant list used in each expression;
- *NbrList* - Number constant list used in each expression;
- *FncList* - Function list used in each expression;
- *IdCount* - Column identifiers count in each expression;
- *StrCount* - String constant count in each expression;
- *NbrCount* - Number constant count in each expression;
- *FncCount* - Functions count in each expression;
- *PrdCount*- Predicate operators count in each expression.

To add additional quantitative measures to each column transformation or column usage in the join and filter conditions we evaluate each expression and calculate transformation and filter weights for those.

Definition 1. Column transformation weight Wt is based on the similarity of each source column and column transformation expression: the calculated weight expresses the source column transfer rate or strength. The weight is calculated on scale [0,1] where 0 means that the data is not transformed form source (e.g. constant assignment in a query) and 1 means that the source is directly copied to the target (no additional column transformations).

Definition 2. Column filter weight Wf is based on the similarity of each filter column in the filter expression and the calculated weight expresses the column filtering rate or strength. The weight is calculated on scale [0,1] where 0 means that the column is not used in the filter and 1 means that the column is directly used in the filter predicate (no additional expressions).

The general column weight W algorithm in each expression for Wt and Wf components is calculated as a column count ratio over all the expression component

[1] http://goldparser.org

counts (e.g. column count, constant count, function count, predicate count).

$$W = \frac{IdCount}{IdCount + FncCount + StrCount + NbrCount + PrdCount}$$

The counts are normalized using the FncList evaluation over a positive function list (e.g. CAST, ROUND, COALESCE, TRIM etc.). If FncList member is in a positive function list then the normalization function reduces the according component count by 1 to "pay a smaller price" in case the function used does not have a significant impact to column data.

Definition 3. A primitive data transformation operation is a data transformation between a source column X and a target column Y in a transformation set M (mapping or query) having expression similarity weight Wt.

Definition 4. The column X is a filter condition in a transformation set M with the filter weight Wf if the column is part of a JOIN clause or WHERE clause in the queries corresponding to M.

3.2 Query Example

Consider the following four SQL queries as an example of related data transformations from source to target tables.

SQL query Q1 parsed to mapping M1:

```
insert into T4(t4.1, t4.2, t4.3)
select T1.t1.1, coalesce(T1.t1.2, '10'),
T1.t1.3
from T1 join T2 ON T2.t2.1 = T1.t1.2
where T2.t2.2 = '10' and T1.t1.2 = 'A'
```

SQL query Q2 parsed to mapping M2:

```
insert into T5 (t5.1, t5.2)
select T4.t4.1, coalesce(T1.t4.2, '10')
from T4 join T3 on T4.t4.1 = T3.t3.1
where T3.t3.2 = '10' and (T1.t4.2 = '10'
or T1.t4.2 is null)
```

SQL query Q3 parsed to mapping M3:

```
insert into T4(t4.1, t4.2, t4.3)
select T6.t6.1, '20', case when T6.t6.2
= 'B' then 20 else 0 end
from T6 join T7 ON T6.t6.3 = T7.t7.1
where 6.t6.2 = 'B' and T7.t7.2 = 'B'
```

SQL query Q4 parsed to mapping M4:

```
insert into T9 (t9.1, t9.2)
select T4.t4.2, coalesce(T4.t4.3*100,
100) from T4 join T8 ON T8.t8.1 =
T4.t4.1
where T4.t4.2 = '20'
```

The queries are parsed to abstract mappings (M1 .. M4) with all the available source and target tables (T1 .. T9). Each mapping has data transformation elements (m1,1 .. m4.2), joins (j1.1 .. j4.1) and filter conditions (f1.1 .. f4.1) according to the query structure and expressions. All the source and target tables have columns (t1.1 .. t9.2) according the usage in query expressions. Additional transformation key-value constraints and conditions (c1, c2, c3) are extracted from the query expressions when possible. The result of the parsed and processed query text is the directed query graph (Figure 2).

Based on the queries metadata, stored mappings, expressions we can calculate the multiple different graphs for data lineage or impact analysis purposes.

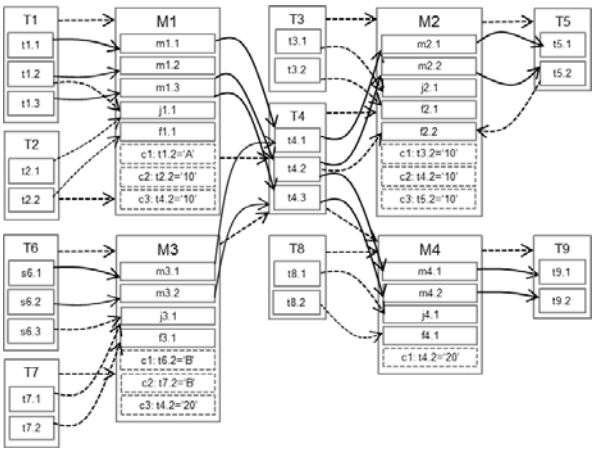


Figure 2 The query transformation graph with all the source and target data structures at column and table level, obtained by parsing and resolving the queries above.

The data lineage graph (Figure 3) illustrates the data flow dependencies and weights calculated by the defined formulas and rules (see in section 4.1). The solid lines stem from the data transformation operations and weights calculated from the query expressions. The dashed lines represent transitive and aggregate relations on the column and the table level.

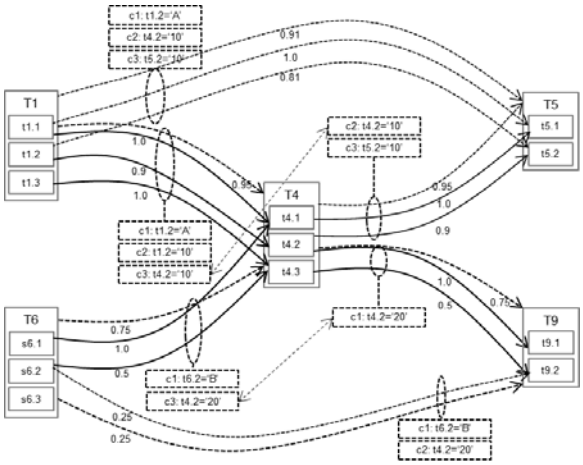


Figure 3 Data lineage graph.

4 Rule System and Dependency Calculation

The source dataset for our rule system based on the database and the queries metadata which capture, parsing and storing is shorty described in previous chapter (steps 1-5 in Fig. 1). The primitive transformations form a graph G_O with nodes N representing database objects (i.e. table or view columns) and edges E_O representing primitive transformations (see Definition 3). We define relations $X: E_O \rightarrow N$ and $Y: E_O \rightarrow N$ connecting edges to source nodes and target nodes, respectively. We define label relations

$M: E_O \rightarrow \{\{m\} \mid m\ is\ a\ transformation\ identifier\}$ and $W: E_O \rightarrow [0,1]$. Formally, this graph is an edge-labeled directed multigraph.

In the remainder of the article, we will use the following intuitive notation: *e.X* and *e.Y* to denote source and target objects of a transformation (formally, $X(e)$ and $Y(e)$). *e.M* is the set of source transformations ($M(e)$). *e.W* is the weight assigned to the edge ($W(e)$).

The knowledge inferred from the primitive transformations forms a graph $G_L = (N, E_L)$ where E_L is the set of edges *e* that represent data flow (lineage). We define relations *X*, *Y*, *M* and *W* the same way as with the graph G_O and use the *e.R* notation where *R* is one of the relations {*X*, *Y*, *M*, *W*}.

Additionally, we designate the graph $G_I = (N, E_I \cup E_L)$ to represent the impact relations between database components. It is a superset of G_L where E_L is the set of additional edges inferred from column usage in filter expressions.

4.1 Rule System

First, we define the rule to map the primitive data transformations to our knowledge base. The result of the rule will be abstract and formalized set of column level dependencies that carry the meaning of data lineage between pair of the source and the target nodes. This rule includes aggregation of multiple edges between pairs of nodes.

Let $E_{x,y} = \{e \in E_O \mid e.X = x, e.Y = y\}$ be the set of edges connecting nodes *x*, *y* in the graph G_O.

$\forall x, y \in N \{E_{x,y} \neq \emptyset \implies \exists e' \in E_L\}$, (R1)

such that

$e'.X = x \wedge e'.Y = y$ (R1.1)

$e'.M = \cup_{e \in E_{x,y}} e.M$ (R1.2)

$e'.W = max\{e.W \mid e \in E_{x,y}\}$ (R1.3)

Inference using this rule should be understood as ensuring that our knowledge base satisfies the rule. From an algorithmic perspective, we create edges *e'* into the set E_L until R1 is satisfied.

Definition 5. The predicate *Parent(x, p)* is true if node *p* is the parent of node *x* in the database schema.

The *Parent* predicate express additional structural dependencies in unified form that we use in next rules (e.g. column's parent is table in original database). Filter conditions are mapped to edges in the impact graph G_I.

Let $F_{M,p} = \{x \mid Parent(x,p) \wedge x\ is\ a\ filter\ in\ M\}$ be the set of nodes that are filter conditions for the mapping *M* with parent *p*. Let $T_{M,p'} = \{x \mid Parent(x,p') \wedge x\ is\ target\ in\ M\}$ be the set of nodes that represent the target columns of mapping *M*.

To assign filter weights to columns, we define the function $W_f: N \rightarrow [0,1]$.

$\forall p, p' \in N \{F_{M,p} \neq \emptyset \wedge T_{M,p'} \neq \emptyset \implies \exists e' \in E_I\}$ (R2),

such that

$e'.X = p \wedge e'.Y = p'$ (R2.1)

$e'.M = M$ (R2.2)

$e'.W = \frac{max\{W_f(x) \mid x \in F_{M,p}\} + max\{W_f(x) \mid x \in T_{M,p'}\}}{2}$ (R2.3)

The primitive transformations mostly represent column-level (or equivalent) objects that are adjacent in the graph (meaning, they appear in the same transformation or query and we have captured the data flow from one to another). The same applies to impact information inferred from filter conditions. From this knowledge, the goal is to additionally:

- propagate information through the database structure upwards, to view data flows on a more abstract level (such as, table or schema level)
- calculate the dependency closure to answer lineage queries

Unless otherwise stated, we treat the the graphs G_L and G_I similarly from this point and it is implied that the described computations are performed on both of them. The set *E* refers to the edges of either of those graphs.

Let $E_{p,p'} = \{e \in E \mid Parent(e.X, p) \wedge Parent(e.Y, p')\}$ be the set of edges where the source nodes share a common parent *p* and the target nodes share a common parent *p'*.

$\forall p, p' \in N \{E_{p,p'} \neq \emptyset \implies \exists e' \in E\}$, (R3)

such that

$e'.X = p \wedge e'.Y = p'$ (R3.1)

$e'.M = \cup_{e \in E_{p,p'}} e.M$ (R3.2)

$e'.W = \frac{\sum_{e \in E_{p,p'}} e.W}{|E_{p,p'}|}$ (R3.3)

4.2 Dependency Closure

Online queries from the dataset require finding the data lineage or impacted set of a database item without long computation times. For displaying both the lineage and impact information, we require that all paths through the directed graph that include a selected component are found. These paths form a connected subgraph. Further manipulation (see Section 4.3) and data display is then performed on this subgraph.

There are two principal techniques for retrieving paths through a node [8]

- connect the edges recursively, forming the paths at query time. This has no additional storage requirements, but is computationally expensive
- store the paths in materialized form. The paths can then be retrieved without recursion, which speeds up the queries, but the materialized transitive closure may be expensive to store.

Several compromise solutions that seek to both efficiently store and query the data have been published [8] and [2]. In general, the transitive closure is stored in a space efficient encoding that can be expanded quickly at query time.

We have incorporated elements from the pointer based technique introduced by Anand et al. [2] The paths are stored in three relations: `Node(N1,P_dep,P_depc)`, `Dep(P_dep,N2)` and `DepC(P_depc,P_dep)`. Immediate dependen-cies of a node are stored in the `Dep` relation, with the pointer `P_dep` in the `Node` relation referring to the dependency set. The full transitive dependency closure is stored in the `DepC` relation by storing the union of the pointers to all

of the immediate dependency sets of nodes along the paths leading to a selected node.

We can define the dependency closure recursively as follows. Let D^*_k be the dependency closure of node k. Let D_k be the set of immediate dependencies such that $D_k = \{j \mid e \in E, e.X = j, e.Y = k\}$.
If $D_k = \emptyset$ then $D^*_k = \emptyset$.
Else if $D_k \neq \emptyset$ then $D^*_k = D_k \cup (\cup_{j \in D_k} D^*_j)$.

The successors S_j (including non-immediate) of a node j are found as follows: $S_i = \{k \mid j \in D^*_k\}$

The materialized storage of the dependency closure allows building the successor set cheaply, so it does not need to be stored in advance. Together with the dependency closure they form the connected maximal subgraph that includes the selected node.

We put the emphasis on fast computation of the dependency closure with the requirement that the lineage graph is sparse ($|E| \sim |N|$). We have omitted the more time-consuming redundant subset and subsequence detection techniques of Anand et al. [1]. The subset reduction has $O(|D|^3)$ time complexity which is prohibitively expensive if the number of initial unique dependency sets $|D|$ is on the order of 10^5 as is the case in our real world dataset.

The dependency closure is computed by:

1. Creating a partial order L of the nodes in the directed graph G_I. If the graph is cyclic then we need to transform it to a DAG by deleting an edge from each cycle. This approach is viable, if the graph contains relatively few cycles. The information lost by deleting the edges can be restored at a later stage, but this process is more expensive than computing the closure on a DAG.
2. Creating the immediate dependency sets for each node using the duplicate-set reduction algorithm of Anand et al. [1].
3. Building the dependency closures for each node using the partial order L, ensuring that the dependency sets are available when they are needed for inclusion in the dependency closures of successor nodes (Algorithm 1)
4. If needed, restoring deleted cyclic edges and incrementally adding dependencies that are carried by those edges using breadth-first search in the direction of the edges.

Algorithm 1. Building the pointer-encoded dependency closure

```
Input: L - partial order on G_I;
 {D_k | k ∈ N} - immediate dependency sets
Output: D*_k - dependency closures for each
node k ∈ N
for node k in L:
    D*_k = {D_k}
    for j in D_k:
        D*_k = D*_k ∪ D*_j
```

This algorithm has linear time complexity $O(|N| + |E|)$ if we disregard the duplicate set reduction. To reduce the required storage, if $D^*_i = D^*_k$ for any $j \neq k$ then we may replace one of them with a pointer to the other. The set comparison increases the worst case time complexity to $O(|N|^2)$.

To extract the nodes along the paths that go through a selected node N, one would use the following queries:

```
select Dep.N2 -- all predecessor nodes
from Node, DepC, Dep
where Dep.P_dep = DepC.P_dep
and DepC.P_depc = Node.P_depc
and Node.N1 = N

select Node.N1 -- all successor nodes
from Node, DepC, Dep
where Node.P_depc = DepC.P_depc
and DepC.P_dep = Dep.P_dep
and Dep.N2 = N
```

4.3 Visualization of the Lineage and Impact Graphs

The visualization of the connected subgraph corresponding to node j is created by fetching the path nodes $P_i = D^*_i \cup S_i$ and the edges along those paths $E_j = \{e \in E \mid e.X \in P_j \wedge e.Y \in P_j\}$ from the appropriate dependency graph (impact or lineage). The graphical representation allows filtering a subset of nodes (in the application, by node type, although the filtering technique discussed here is generic and permits arbitrary criteria). Any nodes not included in graphical display are replaced by transitive edges bypassing these nodes to maintain the connectivity of the dependencies in the displayed graph.

Let $G_j = (P_j, E_j)$ be the connected sub graph for the selected node j. We find the partial transitive graph G_j' that excludes the filtered nodes P_{filt} as follows (Algorithm 2):

Algorithm 2. Building the filtered subgraph with transitive edges

```
Input: G_j, P_filt
Output: G_j' = (P_j', E_j')
E_j' = E_j
P_j' = ∅
for node n in P_j:
 if n ∈ P_filt:
   for e in {e ∈ E_j' | e.Y = n}:
     for e' in {e' ∈ E_j' | e'.X = n}:
       create new edge e'' ( e''.X = e.X, e''.Y
= e'.Y, e''.W = e.W * e''.W)
       E_j' = E_j' ∪ {e''}
     E_j' = E_j' \ {e}
   for e' in {e' ∈ E_j' | e'.X = n}:
     E_j' = E_j' \ {e'}
 else:
   P_j' = P_j' ∪ {n}
```

This algorithm has the time complexity of O(|Pj| + |Ej|) and can be performed on demand when the user changes the filter settings. This extends to large dependency graphs with the assumption that |GJ| << |G|.

4.4 Semantic Layer Calculation

The semantic layer is additional visualization and specific filter set to localize connected subgraph of the expected data flows for the current selected node. All connected nodes and edges in the semantic layer share

the overlapping filter predicate conditions or data production conditions that are extracted during the edge construction to indicate not only possible data flows (based on connections in initial query graph), but only expected and probabilistic data flows. Main idea of the semantic layer is to narrow down all possible and expected data flows over all connected graph nodes by cutting down unlikely of not allowed connections in graph, which based on additional query filters and semantic interpretation of filters and calculated transformation expression weights. Semantic layer of data lineage graph will hide irrelevant or highlights relevant graph nodes and edges (depends on user choice and interaction) that makes huge distinction when underlying data structures are abstract enough and independent data flows store and use independent "horizontal" slices of data. The essence of semantic layer is to use available query and schema information to estimate the row level data flows without additional row level lineage information that is unavailable on schema level, but also expensive or impossible to collect on row level.

The visualization of the semantically connected subgraph corresponding to node j is created by fetching the path nodes $P_j = D^*_j \cup S_j$ and the edges along those paths $E_j = \{e \in E | e.X \in P_j \wedge e.Y \in P_j\}$ from the appropriate dependency graph (impact or lineage). Any nodes not included in semantic layer are removed or visually muted (by change color or opacity) and semantically connected subgraph returned or visualized in UI.

Let $G_j = (P_j, E_j)$ be the connected subgraph for the selected node j where $GD_j = (D_j, ED_j)$ is predecessor subgraph and $GS_j = (S_j, ES_j)$ is successor subgraph according to selected node *j*. We find the data flow graph G_j' that is union of semantically connected predecessors $GD_j' = (D_j, ED_j)$ and successor subgraphs $GS_j' = (S_j, ES_j)$. The semantic layer calculation based on selected node filter set F_j and calculated separately for back (predecessor) and forward (successors) direction by similar recursive algorithm (Algorithm 3):

Algorithm 3. Building the semantic layer subgraph using predecessors and successor recursive functions

```
Function: Predecessors
  Input: nj, Fj , GDj, GD'j Wmin
  Output: GDj' = (Dj', EDj')
  Fn = ∅
  if Dj' = ∅ then:
   Dj' = Dj' ∪ nj
  for edge e in {e ∈ EDj | e.Y = nj
   Fn = ∅
   if Fj != ∅:
     for filter f in e.{F}:
       for filter fj in Fj:
         if f.Key = fj.Key & f.Val ∩ fj.Val:
             new   filter   fn   (fn.Key=f.Key,
  fn.Val=f.Val, fn.Wgt=f.Wgt*fj.Wgt)
           Fn = Fn ∪ fn
   else:
     Fn = Fn ∪ e.{F}
   if Fn != ∅ & e.W >= Wmin :
     Dj' = Dj' ∪ e.X
     EDj' = EDj' ∪ e
     GDj' = Predecessors(e.X,Fn,GDj,GD'j,Wmin)
  return GDj'

Function: Successors
  Input: nj, Fj , GSj, GS', Wmin
  Output: GSj' = (Sj', ESj')
  Fn = ∅
  if Sj' = ∅:
   Sj' = Sj' ∪ nj
  for edge e in {e ∈ ESj | e.X = nj}:
   Fn = ∅
   if Fj != ∅ then:
     for filter f in e.{F}:
       for filter fj in Fj:
         if f.Key = fj.Key & f.Val ∩ fj.Val:
           new   filter   fn   (fn.Key=f.Key,
  fn.Val=f.Val, fn.Wgt=f.Wgt*fj.Wgt)
           Fn = Fn ∪ fn
   else:
     Fn = Fn ∪ e.{F}
   if Fn != ∅ & e.W >= Wmin :
     Sj' = Sj' ∪ e.Y
     ESj' = ESj' ∪ e
     GSj' = Predecessors(e.Y,Fn ,GSj,GSj',Wmin)
  return GSj'
```

The final semantic layer subgraph is union of recursively constructed predecessors GD_j' and successor GS_j' graphs: `Gj' = GDj' ∪ GSj'`

5 Case Studies

The previously described architecture (see Figure 1 in Section 3) rule system and and algorithms (in Section 4) have been used to implement an integrated toolset dLineage[2]. The toolset is built as an independent application or web based service that collect DW metadata, store and process it in internal PostgreSQL database, and serve the processed data and calculated graph results as interactive multilayer map for further analysis. Both the scanners and web-based tools of dLineage have been enhanced and tested in real-life projects and environments to support several popular DW database platforms (e.g. Oracle, Greenplum, Teradata, Vertica, PostgreSQL, MsSQL, Sybase), ETL tools (e.g. Informatica, Pentaho, Oracle Data Integrator, SSIS, SQL scripts and different data loading utilities) and BI tools (e.g. SAP Business Objects, Microstrategy, SSRS). The dLineage dynamic visualization and graph navigation tools are implemented in Javascript using the d3.js graphics libraries.

Current implementation has rule system which is implemented in PostgreSQL database using SQL queries for graph calculation (rules 1-3 in section 4.1) and specialized tables for graph storage. The DB and UI interaction tested with the specialized pre-calculated model (see section 4.2) but also with the recursive queries without special storage and calculations. The algorithms for interactive transitive calculations (see sections 4.3) and semantic layer calculation (see section 4.4) are implemented in Javascript and works in browser for small and local subgraph optimization or

[2] http://dlineage.com

visualization. Due to space limitations we do not stop here for performance figures and discussion, but just illustrate the application and algorithms with the visualizations in next chapter.

5.1 Dataset Visualization

The Enterprise Dependency Graph examples (Figures 4-6) are illustrations of the complex structure of dependencies between the DW storage scheme, access views and user reports. The example is generated using DW and BI lineage layers and has details at database and reporting object level (not at column level). At the column and report level the full data lineage graph would be about ten times bigger and too complex to visualize in a single picture. The following graph from DW structures and user reports presents about 50,000 nodes (tables, views, scripts, queries, reports) and 200,000 links (data transformations in views and queries) on a single image (Figure 4).

The real-life dependency graph examples illustrate the automated data collection, parsing, resolving, graph calculation and visualization tasks implemented in our system. The system requires only the setup and configuration tasks to be performed manually. The rest will be done by the scanners, parsers and the calculation engine.

The end result consists of data flows and system component dependencies visualized in the navigable and drillable graph or table form. The result can be viewed as a local subgraph with fixed focus and suitable filter set to visualize data lineage path from any sources to single report with click and zoom navigation features. The big picture of the dependency network gives the full scale overview graph of the organization's data flows. It allows to see us possible architectural, performance or security problems.

6 Conclusions and Future Work

We have presented several algorithms and techniques for quantitative impact analysis, data lineage and change management. The focus of these methods is on automated analysis of the semantics of data conversion systems followed by employing probabilistic rules for calculating chains and sums of impact estimations. The algorithms and techniques have been successfully employed in several large case studies, leading to practical data lineage and component dependency visualizations. We continue this research by performance measurement with the number of different big datasets, to present practical examples and draw conclusion of our approach.

We also considering a more abstract, conceptual and business level approach in addition to the current physical/technical level of data lineage representation and automation.

Figure 4 Data flows (blue,red) and control flows (green,yellow) between DW tables, views and reports.

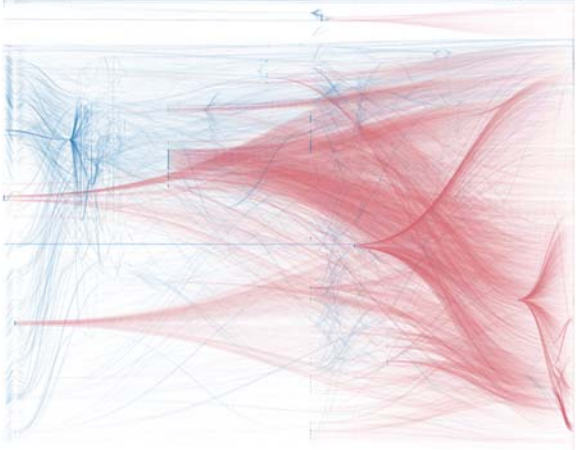


Figure 5 Data flows between DW tables, views (blue) and reports (red).

Figure 6 Control flows in scripts, queries (green) and reporting queries (yellow) are connecting DW tables, views and reports.

Acknowledgments

The research has been supported by EU through European Regional Development Fund.

References

[1] Anand, M. K., Bowers, S., McPhillips, T., & Ludäscher, B. (2009, March). Efficient provenance storage over nested data collections. In *Proceedings of the 12th International Conference on Extending Database Technology: Advances in Database Technology* (pp. 958-969). ACM.

[2] Anand, M. K., Bowers, S., & Ludäscher, B. (2010, March). Techniques for efficiently querying scientific workflow provenance graphs. In *EDBT* (Vol. 10, pp. 287-298).

[3] Cui, Y., Widom, J., & Wiener, J. L. (2000). Tracing the lineage of view data in a warehousing environment. ACM Transactions on Database Systems (TODS), 25(2), 179-227.

[4] Cui, Y., & Widom, J. (2003). Lineage tracing for general data warehouse transformations. The VLDB Journal—The International Journal on Very Large Data Bases, 12(1), 41-58.

[5] de Santana, A. S., & de Carvalho Moura, A. M. (2004). Metadata to support transformations and data & metadata lineage in a warehousing environment. In Data Warehousing and Knowledge Discovery (pp. 249-258). Springer Berlin Heidelberg.

[6] Fan, H., & Poulovassilis, A. (2003, November). Using AutoMed metadata in data warehousing environments. In Proceedings of the 6th ACM international workshop on Data warehousing and OLAP (pp. 86-93). ACM.

[7] Giorgini, P., Rizzi, S., & Garzetti, M. (2008). GRAnD: A goal-oriented approach to requirement analysis in data warehouses. Decision Support Systems, 45(1), 4-21.

[8] Heinis, T., & Alonso, G. (2008, June). Efficient lineage tracking for scientific workflows. In *Proceedings of the 2008 ACM SIGMOD international conference on Management of data* (pp. 1007-1018). ACM.

[9] Ikeda, R., Das Sarma, A., & Widom, J. (2013, April). Logical provenance in data-oriented workflows?. In *Data Engineering (ICDE), 2013 IEEE 29th International Conference on* (pp. 877-888). IEEE.

[10] Missier, P., Belhajjame, K., Zhao, J., Roos, M., & Goble, C. (2008). Data lineage model for Taverna workflows with lightweight annotation requirements. In Provenance and Annotation of Data and Processes (pp. 17-30). Springer Berlin Heidelberg.

[11] Priebe, T., Reisser, A., & Hoang, D. T. A. (2011). Reinventing the Wheel?! Why Harmonization and Reuse Fail in Complex Data Warehouse Environments and a Proposed Solution to the Problem.

[12] Ramesh, B., & Jarke, M. (2001). Toward reference models for requirements traceability. Software Engineering, IEEE Transactions on, 27(1), 58-93.

[13] Reisser, A., & Priebe, T. (2009, August). Utilizing Semantic Web Technologies for Efficient Data Lineage and Impact Analyses in Data Warehouse Environments. In Database and Expert Systems Application, 2009. DEXA'09. 20th International Workshop on (pp. 59-63). IEEE.

[14] Skoutas, D., & Simitsis, A. (2007). Ontology-based conceptual design of ETL processes for both structured and semi-structured data. International Journal on Semantic Web and Information Systems (IJSWIS), 3(4), 1-24.

[15] Tomingas, K., Tammet, T., & Kliimask, M. (2014), Rule-Based Impact Analysis for Enterprise Business Intelligence. In Proceedings of the Artificial Intelligence Applications and Innovations (AIAI2014) conference workshop (MT4BD). Series: IFIP Advances in Information and Communication Technology, Vol. 437.

[16] Tomingas, K., Kliimask, M., & Tammet, T. (2015). Data Integration Patterns for Data Warehouse Automation. In New Trends in Database and Information Systems II (pp. 41-55). Springer International Publishing.

[17] Vassiliadis, P., Simitsis, A., & Skiadopoulos, S. (2002). Conceptual modeling for ETL processes. In Proceedings of the 5th ACM international workshop on Data Warehousing and OLAP (pp. 14-21). ACM.

[18] Widom, J. (2004). Trio: A system for integrated management of data, accuracy, and lineage. Technical Report.

[19] Woodruff, A., & Stonebraker, M. (1997). Supporting fine-grained data lineage in a database visualization environment. In Data Engineering, 1997. Proceedings. 13th International Conference on (pp. 91-102). IEEE.

Conceptual Modeling with Formal Concept Analysis on Natural Language Texts †

© Mikhail Bogatyrev
Tula State University,
Tula
okkambo@mail.ru

Abstract

The paper presents conceptual modelling technique on natural language texts. This technique combines the usage of two conceptual modeling paradigms: conceptual graphs and Formal Concept Analysis. Conceptual graphs serve as semantic models of text sentences and the data source for concept lattice – the basic conceptual model in Formal Concept Analysis. With the use of conceptual graphs the Text Mining problems of Named Entity Recognition and Relations Extraction are solved. Then these solutions are applied for creating concept lattice. The main problem investigated in the paper is the problem of creating formal contexts on a set of conceptual graphs. Its solution is based on the analysis of semantic roles and conceptual patterns in conceptual graphs. Concept lattice built on textual data is applied for knowledge extraction. Knowledge, sometimes interpreted as facts, can be extracted by using navigation in the lattice and interpretation its concepts and hierarchical links between them. Experimental investigation of the proposed technique is performed on the annotated textual corpus consisted of descriptions of biotopes of bacteria.

†The paper concerns the work which is partially supported by Russian Foundation of Basic Research, grant № 15-07-05507

1 Introduction

Knowledge extraction from textual data requires more in-depth intensive analysis of this data. In the area of Text Mining, some variants of knowledge extraction have been realized by solving such problems as *sentiment analysis*, *fact extraction* and *decision making support*. To solve these problems it is necessary to have models that reflect semantics of textual data. It is especially urgent when this data is presented as unstructured natural language texts.

Proceedings of the XVIII International Conference «Data Analytics and Management in Data Intensive Domains» (DAMDID/RCDL'2016), Ershovo, Russia, October 11 - 14, 2016

Conceptual modeling is one of the ways of modeling semantics in the Natural Language Processing (NLP) [22]. Conceptual modeling is the process of conceptualization of real world phenomena and creating conceptual models as a result of conceptualization. Conceptual model is a graph which vertices are concepts and arrows or edges are links between concepts. Every conceptual model has its own semantics which represents the meanings of concepts and links.

Conceptual modeling has long been applied for databases and software modeling [19] and this term is also used in other fields including NLP. Entity Relationship Diagram (ERD) [19] is well known representative of conceptual models. It describes the structure of database in terms of *entities*, *relationships*, and *constraints*. These terms of entities, relationships, and constraints are explicitly or implicitly present at many other conceptual models including ones discussed in this paper.

Formal Concept Analysis (FCA) [13] is the paradigm of conceptual modeling which studies how objects can be hierarchically grouped together according to their common attributes. In the FCA, its conceptual model is the lattice of formal concepts (concept lattice) which is built on the abstract sets treated as objects and their attributes. Concept lattices have been applied as an instrument for information retrieval and knowledge extraction in many applications. The number of FCA applications now is growing up including applications in social science, civil engineering, planning, biology, psychology and linguistics [21], [22]. Several successful implementations of FCA methods on fact extraction on textual data [8] and Web data are known [15]. Although the high level of abstraction makes FCA suitable for use with data of any nature, its application to specific data often requires special investigation. It is fully relevant for using FCA on textual data.

The main problem in creating concept lattice on textual data is building so called *formal contexts* on this data. Formal context is matrix representation of the relation on the sets of objects and attributes. So it is needed to acquire words or word combinations from texts which are interpreted as objects and attributes. To restrict all possible combinations of words of such meanings we need to select from them those ones which are valued for solving concrete problem or the class of problems. As a result a concept lattice created on texts

becomes domain specific. This is similar to the design of ontologies and concept lattice is often considered as framework of ontology [21].

Another paradigm of conceptual modeling is Conceptual Graphs (CGs) [24]. Conceptual graph is bipartite directed graph having two types of vertices: concepts and conceptual relations. Conceptual terms of entities and relationships are represented in conceptual graphs as its concepts and conceptual relations.

Conceptual graphs have been applied for modeling many real life objects including texts. Acquiring conceptual graphs from natural language texts is non-trivial problem but it is quite solvable [3], [5].

The main purpose of this paper is to show how two paradigms of conceptual modeling - Conceptual Graphs and Formal Concept Analysis - can be united in one modeling technique. The idea of joining these two paradigms seems very attractive but not elaborated much enough [22], [26].

Proposed technique is used in on-going project of creating fact extraction system working on biomedical data. Experimental investigation of it is performed on the annotated textual corpus consisted of descriptions of biotopes of bacteria.

2 CGs-FCA modeling

The proposed modeling technique named briefly as CGs-FCA modeling is based on using conceptual graphs and concept lattice. It may be applied for knowledge extraction from textual data. In CGs-FCA modeling conceptual graphs serve as semantic models of text sentences and the data source for formal context of concept lattice. Concept lattice built on textual data is applied for knowledge extraction. Knowledge, sometimes interpreted as facts, can be extracted by using navigation in the lattice and interpretation its concepts and hierarchical links between them.

To illustrate CGs-FCA modeling, consider some FCA basics.

2.1 Formal Concept Analysis basics

There are two basic notions FCA deals with: *formal context* and *concept lattice* [13]. Formal context is a triple $\mathbf{K} = (G, M, I)$, where G is a set of objects, M – set of their attributes, $I \subseteq G \times M$ – binary relation which represents facts of belonging attributes to objects. The sets G and M are partially ordered by relations ϕ and P, correspondingly: $G = (G, \phi)$, $M = (M, \mathrm{P})$. Formal context may be represented by [0, 1] - matrix $\mathbf{K} = \{k_{i,j}\}$ in which units mark correspondence between objects $g_i \in G$ and attributes $m_j \in M$. The concepts in the formal context have been determined by the following way. If for subsets of objects $A \subseteq G$ and attributes $B \subseteq M$ there are exist mappings (which may be functions also) $A' : A \to B$ and $B' : B \to A$ with properties of $A' := \{\exists m \in M \mid < g, m > \in I \ \forall g \in A\}$ and $B' := \{\exists g \in G \mid < g, m > \in I \ \forall m \in B\}$ then the pair (A, B) that $A' = B$, $B' = A$ is named as *formal concept*. The sets A and B are closed by composition of mappings: $A'' = A$, $B'' = B$; A and B is called the *extent* and the *intent* of a formal context $\mathbf{K} = (G, M, I)$ respectively.

By other words, a formal concept is a pair (A, B) of subsets of objects and attributes which are connected so that every object in A has every attribute in B, for every object in G that is not in A, there is an attribute in B that the object does not have and for every attribute in M that is not in B, there is an object in A that does not have that attribute.

The partial orders established by relations ϕ and P on the set G and M induce a partial order $\leq$ on the set of formal concepts. If for formal concepts (A_1, B_1) and (A_2, B_2), $A_1 \, \phi \, A_2$ and $B_2 \, \mathrm{P} \, B_1$ then $(A_1, B_1) \leq (A_2, B_2)$ and formal concept (A_1, B_1) is less general than (A_2, B_2). This order is represented by *concept lattice.* A lattice consists of a partially ordered set in which every two elements have a unique *supremum* (also called a least upper bound or *join*) and a unique *infimum* (also called a greatest lower bound or *meet*).

According to the central theorem of FCA [13], a collection of all formal concepts in the context $\mathbf{K} = (G, M, I)$ with subconcept-superconcept ordering $\leq$ constitutes the *concept lattice* of $\mathbf{K}$. Its concepts are subsets of objects and attributes connected each other by mappings A', B' and ordered by a subconcept-superconcept relation.

A	B	C	D	E
	Membrane	Nucleus	Replication	Recombination
DNA				X
Virus	X			X
Prokaryotes	X		X	
Eukaryotes	X	X	X	
Bacterium	X		X	

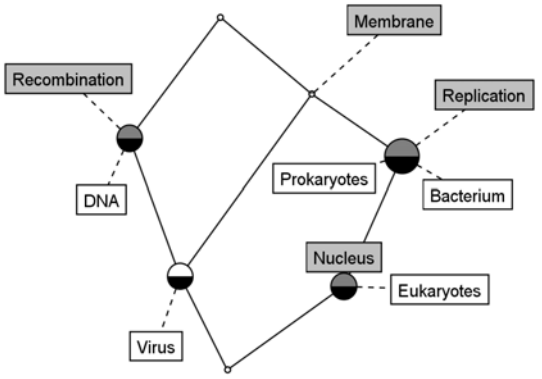


Figure 1 Example of formal context and concept lattice.

To illustrate these abstract definitions consider an example. Figure 1 shows simple formal context and concept lattice composed on the sets ***G*** = {*DNA, Virus, Prokaryotes, Eukaryotes, Bacterium*} and ***M*** = {*Membrane, Nucleus, Replication, Recombination*}. The

set ***G*** is ordered according to sizes of its elements: DNA is smallest and bacterium is biggest ones. The set ***M*** has relative order: one part (*Membrane, Nucleus*) characterizes microbiological structure of objects from ***G***, but another part (*Replication, Recombination*) characterizes the way of breeding, and these parts are incomparable. In the concept lattice the bacterium is placed in the concept C_1 = ({*Prokaryotes, Eukaryotes, Bacterium*}, {*Membrane, Replication*}). In this concept, three objects {*Prokaryotes, Eukaryotes, Bacterium*} constitute the extent of the concept; they are united by their mutual attribute {*Membrane, Replication*} which constitute the intent of the concept. The concept C_1 is more general concept than the concept C_2 = ({*Eukaryotes*}, {*Nucleus*}).

Also on the Fig. 1 there are two different branches of concepts characterizing two families: the viruses and DNA and prokaryotes, eukaryotes and bacteria. This concept demonstrates the fact of separation of objects from the set *G* into two important branches. The link between them is the attribute "*Membrane*". It is known [7] that viruses can have a lipid shell formed from the membrane of the host cell. Therefore, the membrane is positioned in the formal context on the Fig. 1 as an attribute of the virus.

This example demonstrates specific ways of extracting knowledge from conceptual lattice:

- analyzing formal concepts in concept lattice;
- analyzing conceptual structures in concept lattice – its sub lattices in the general case.

2.1.1 FCA on textual data

The main problem in applying FCA on textual data is the problem of building formal context. If textual data is represented as natural language texts then this problem becomes acute.

There are several approaches to the construction of formal contexts on the textual data, presented as separate documents, as data corpora. One, mostly applied variant is the context in which the objects are text documents and the attributes are the terms from these documents. Another variant is building formal context directly on the texts and the formal context may represent various features of textual data:

- semantic relations (synonymy, hyponymy, hypernymy) in a set of words for semantic matching [16],
- verb-object dependencies from texts [10],
- words and their lexico-syntactic contexts [20].

These lexical elements must be distinguished in texts as objects and attributes. There are following approaches to solve this problem:

- adding special descriptions to texts which mark objects and attributes and partial order,
- using corpus tagging and semantic models of texts [10].

We apply the second approach and use conceptual graphs for representing semantics of individual sentences of a text.

2.2 The modeling process

Consider in general the process of CGs – FCA modeling. It includes the following steps.

1. *Acquiring a set of conceptual graphs from input texts*. As it is mentioned above conceptual graphs can be acquired from texts by existing information systems. For example they can be created by our system CGs Maker [1]. Some details about it can be found in [3], [5]. We use verb-centered approach for creating conceptual graphs. According to this approach, a conceptual graph is constructed so that there is the central concept in it which is realized as a verb. If there are no verbs in a sentence then method also creates conceptual graph. Verb-centered approach is important for us since it provides predicate forms in the structures of conceptual graphs. These forms are mostly used for representing conceptual graph semantics.

2. *Aggregating the set of conceptual graphs*. Aggregation is needed to exclude excessive dimension of conceptual models, not related to useful information. We have tested following ways of conceptual graphs aggregation: conceptual graphs clustering and using corpus tagging together with support of concept types in conceptual graphs. Clusters of conceptual graphs need to be semantically interpreted which may lead to additional investigations. The second method is more constructive since it selects those conceptual graphs which concepts have mappings to certain domain. Such domain of terms may be presented by corpus tagging or by thesaurus. Some details of aggregation are below.

3. *Creating formal contexts*. This is the central point of CGs – FCA modeling. One or several formal contexts have been built on the aggregated conceptual graphs. The number of formal concepts and the method of building them depend on the problem being solved with CGs – FCA modeling.

4. *Building concept lattice*. Having a formal context as input data, a concept lattice may be created by using various algorithms. There is a field of research in FCA devoted to creating and developing algorithms for concept lattice creation [21]. On the current stage of CGs – FCA modeling technique we use standard solution of creating concept lattice realized in the open source tool [27]. Nevertheless, here there are certain possibilities to create new algorithms oriented on specific structure of formal contexts acquired from conceptual graphs. One of such structure is block-diagonal structure which arises namely on using textual data as input.

5. *Knowledge extraction from concept lattice*. In concept lattice it is possible to identify connections

[1] The lightweight online version of CGs Maker for simple English and Russian texts can be found at http://85.142.138.156:8888 .

between its concepts according to the principle of "common – particular". Each concept may be interpreted as the set of potential facts of certain level, which is associated with other facts. So the knowledge extracted from concept lattice may be interpreted a s facts.

2.3 Aggregation of conceptual graphs

In the theory of conceptual graphs aggregation means replacing conceptual graphs by more general graphs [24] . These general graphs may be created as new graphs or may be graphs or sub graphs from initial set of graphs. Aggregation of conceptual graphs has semantic meaning and general graphs make up *the context* (not formal context) of initial set of graphs.

Clustering is a way of aggregation of conceptual graphs. Graphs which are the nearest ones to the centers of clusters have been treated as general graphs.

We have studied several approaches for clustering conceptual graphs [2] using various similarity measures. There are two known similarity measures proposed in [17], the conceptual similarity

$$s_c = \frac{2n(\gamma_c)}{n(\gamma_1)+n(\gamma_2)} \quad (1)$$

and relative similarity

$$s_r = \frac{2m(\gamma_c)}{m_{\gamma_c}(\gamma_1)+m_{\gamma_c}(\gamma_2)} \quad (2)$$

Here γ_1, γ_2 - conceptual graphs, $\gamma_c = \gamma_1 \cap \gamma_2$ is their common sub graph, $n(\gamma_i)$ - number of concepts of graph γ_i, $m(\gamma_i)$ - number of relations of graph γ_i, $m_{\gamma_c}(\gamma_i)$ is the number of relations of conceptual graph γ_i, at least one of which belongs to the common sub graph γ_c.

If two conceptual graphs have identical concepts then their conceptual similarity has non zero value. Relative similarity is non-zero when two conceptual graphs have identical structures of patterns of conceptual relations.

We used conceptual and relative similarities (1), (2) and their combination in the experiments of conceptual graphs clustering [2]. Except traditional algorithms of clustering such as *K*-means, we used genetic clustering algorithm with special encoding. The peculiarity of implementing genetic algorithms for clustering is that there may be several final solutions i.e. several different variants of clustering.

All numerical characteristics of conceptual graphs clustering results (number of clusters, dimensions of clusters, etc.) are not informative. Clusters of conceptual graphs need to be semantically interpreted. The way of that interpretation depends on the nature of the problem to be solved with conceptual graphs.

Both conceptual and relative similarity measures share a common sub graph γ_c. But two conceptual graphs may have no common sub graph but may be similar "semantically". That means that their concepts have the same type. For example different names of bacteria belong to the type "bacterium" or the type "the name of bacteria".

The second way of conceptual graphs aggregation is based on supporting types of concepts by using external resources. Thesaurus or corpus tagging may be such resource. Section 3 contains additional details.

2.4 Creating formal contexts

The crucial step in the described process of CGs – FCA modeling is creating formal contexts on the set of conceptual graphs.

At first glance, this problem seems simple: those concepts of conceptual graphs which are connected by "attribute" relation have been put into formal context as its objects and attributes. Actually the solution is much more complex.

Fig. 2 shows an example of conceptual graph for the sentence "*Burkholderia phytofirmans belongs to the beta-proteobacteria and was isolated from surface-sterilized glomus vesiculiferum-infected onion roots.*"

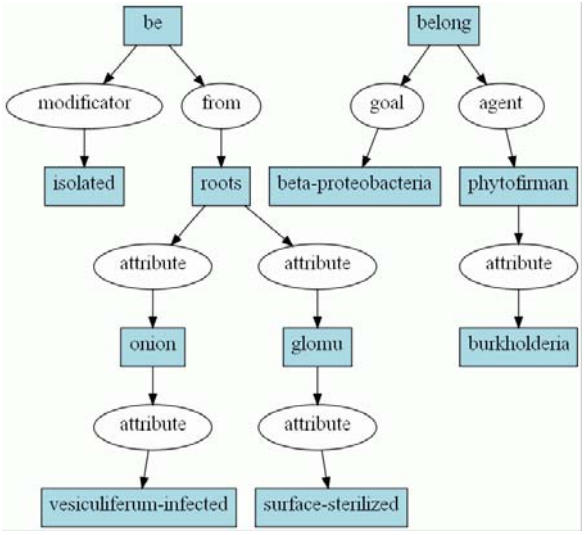


Figure 2 Conceptual graph for the sentence "*Burkholderia phytofirmans belongs to the beta-proteobacteria and was isolated from surface-sterilized glomus vesiculiferum-infected onion roots.*"

This graph has five conceptual relations "attribute" but only four of them indicate the objects and attributes valid for formal context. Using "*phytofirman*" as object and "*Burkholderia*" as attribute in the formal context is wrong way because "*Burkholderia phytofirmans*" is known full name of this bacterium [6] and full names of bacteria have to be objects in a formal context devoted to bacteria. Word combinations denoting the names of bacteria must be recognized before the building of conceptual graphs. There is no other way of doing this than to use an external source of information, for example, the corpus tagging. So in this example the sub

graph <*phytofirman*>- (*attribute*) - < *burkholderia* > is useless for bacteria names recognizing.

Remaining elements of conceptual graph on the Fig. 2 are not useless and play significant roles in creating formal context. Conceptual graph on the Fig. 2 represents two facts:

1. bacterium *Burkholderia phytofirmans* belongs to beta-proteobacteria;
2. this bacterium infects the onion.

To provide the presence information about those and other facts in the formal contexts the following rules are implemented as mostly important when creating formal contexts.

1. Not only individual concepts and relations, but also patterns of connections between concepts in conceptual graphs represented as sub graphs have been analyzed and processed. These patterns are predicate forms <object> - <predicate> - <subject> which in conceptual graphs look as the template <concept>- (patient) - < verb > - (agent) - <concept>. Not only agent and patient semantic roles but also other similar to them (goal on the Fig. 2) roles are allowed in templates.
2. The hierarchy of conceptual relations in conceptual graphs is fixed and taken into account when creating formal context. This hierarchy exists on the Fig.2: relations “agent”, “goal”, “from”, “modificator” are on the top level and relations "attribute" belong to underlying levels. Using this hierarchy of conceptual relations we can select for formal contexts more or less details from conceptual graphs.

These empirical rules are related to the principle of *pattern structures* which was introduced in FCA in the work [12]. A pattern structure is the set of objects with their descriptions (patterns), not attributes. Patterns also have similarity operation. The instrument of pattern structures is for creating concept lattices on the data being more complicated than sets of objects and attributes.

Conceptual graph is a pattern for the object it represents. A sub graph of conceptual graph is *projection* of a pattern. Namely projections are often used for creating formal contexts. Similarity operation on conceptual graphs is a measure of similarity which is applied in clustering. The relative similarity (2) is mostly close to be similarity operation for patterns.

The CGs – FCA modeling technique was tested in various levels of its realization for classification messages in technical support services [3], modeling requirements for information systems [4] and classifying queries to biomedical systems [5].

3 CGs-FCA modeling on biomedical data

3.1 Biomedical data intensive domain

Bioinformatics is the field where Data Mining and Text Mining applications are growing up rapidly. New term of “Biomedical Natural Language Processing” (BioNLP) has been appeared there [1]. This appearing is stipulated by huge amount of scientific publications in Bioinformatics and organizing them into corpora with access to full texts of articles via such systems as PubMed [25]. Information resources of PubMed have been united in several subsystems presenting databases, corpora and ontologies.

So called “research community around PubMed” [14] forms data intensive domain in this area. It not only uses data from PubMed but also creates new data resources and data mining tools including specialized languages for effective biomedical data processing [11].

In our experiments we use PubMed vocabulary thesaurus MeSH (Medical Subject Headings) as external resource for supporting types of concepts in conceptual graphs.

3.2 Data structures

Our experiments have been carried out using text corpus of bacteria biotopes which is used in the innovation named as BioNLP Shared Task [6]. Biotope is an area of uniform environmental conditions providing a living place for plants, animals or any living organism. Biotope texts form tagged corpus. The tagging includes full names of bacteria, its abbreviated names and unified key codes in the database. We can add additional tags and we do it.

A BioNLP data is always domain-specific. All the texts in the corpus are about bacteria themselves, their areal and pathogenicity. Not every text contains these three topics but if some of them are in the text then they are presented as separate text fragments. This simplifies text processing.

The CGs – FCA modeling environment has DBMS for storing and managing data used in experiments. We use relational database on the SAP-Sybase platform. Database stores texts, conceptual graphs, formal contexts and concept lattices. Special indexing is applied on textual data.

3.3 BioNLP tasks

According to the BioNLP Shared Task initiative [6] there are two main tasks solving on biomedical corpora: the task of Named Entity Recognition (NER) and the task of Relations Extraction (RE).

The task of Named Entity Recognition on the corpus of bacteria descriptions is formulated as seeking bacteria names presented directly in the texts or as co-references (anaphora).

Relations Extraction means seeking links between bacteria and their habitat and probably diseases it causes.

3.4 NER and anaphora resolution

The task of Named Entity Recognition has direct solution with conceptual graphs. The only problem which is here is anaphora resolution.

Anaphora resolution is the problem of resolving references to earlier or later items in the text. These items are usually noun phrases representing objects called referents but can also be verb phrases, whole sentences

or paragraphs. Anaphora resolution is the standard problem in NLP.

Biotopes texts we work with contain several types of anaphora:

- hyperonym definite expressions ("bacterium" - "organism", "cell" - "bacterium"),
- higher level taxa often preceded by a demonstrative determinant ("this bacteria", "this organism"),
- sortal anaphors ("genus", "species", "strain").

For anaphora detection and resolution we use a pattern-based approach. It is based on fixing anaphora items in texts and establishing relations between these items and bacteria names. We use double-pass algorithm for anaphora resolution which controls so called isolated concepts appeared on the first pass of the algorithm. Isolated concepts are those concepts which are not connected by relation with any other concepts. As a rule they appear when a sentence contains abbreviations or code of bacterium. For example, in the sentence "*Streptococcus thermophilus strain LMG 18311*" there is code of bacterium strain. This code will be presented as isolated concept in conceptual graph. Later in another sentence there is text fragment "*…two yogurt strains of S. thermophiles* …" which has abbreviation of the name of bacterium. Having isolated concept with strain code we can identify it with bacterium using corpus tagging. For resolving abbreviations programming triggers which react to the second word after abbreviation are applied.

To evaluate the quality of this solution of NER the standard characteristics of recall, precision and *F*-score were calculated. To obtain them it was needed to mark named entities manually in the texts used in experiments. The table 1 contains values of recall, precision and *F*-score compared with corresponding values from the work [23]. In this work pattern-based approach is also applied and several external resources were involved in the NER solution. The Alvis system was explored in [23] and SemText is the name of our system which explores CGs – FCA modeling.

Table 1 Recall, precision and *F*-score for NER solutions

	Recall	Precision	*F*-score
Alvis	0,52	0,46	0,59
SemText	0,42	0,53	0,47

The ratio of the values of recall and precision is more informative than their individual ones and is shown on the Fig. 3. According to the table 1 and Fig. 3 we resume that there is medium quality of our solution of NER. It is explained by disability of our algorithm to interpret all possible isolated concepts in conceptual graph. As a result approximately half of marked lexical elements were not recognized as entities.

3.5 Relations extraction with concept lattices

Conceptual graphs represent relations between words. Therefore they can be applied for relations extraction but only in one sentence. For extracting relations between bacteria on several texts we applied concept lattices.

We had selected 130 mostly known bacteria and have processed corresponding corpus texts about them. All the texts were preliminary filtered for excluding stop words and other non-informative lexical elements.

Three formal contexts of "Entity", "Areal" and "Pathogenicity" were built on the texts. They have the names of bacteria as objects and corresponding concepts from conceptual graphs as attributes. Table 2 shows numerical characteristics of created contexts.

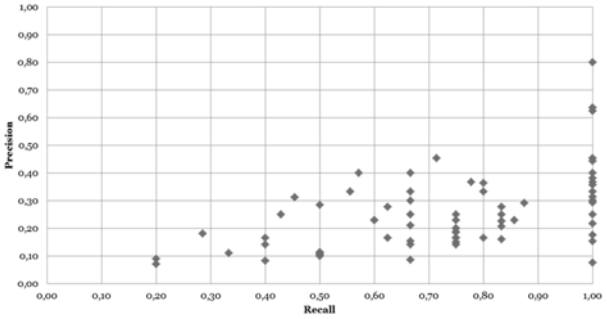


Figure 3 Recall and precision ratio for NER solution on 60 objects

Table 2 Numerical characteristics of created contexts

Context name	Number of objects	Number of attributes	Number of formal concepts
Entity	130	26	426
Areal	130	18	127
Pathogenicity	130	28	692

Among attributes there are bacteria properties (gram-negative, rod-shaped, etc.) for "Entity" context, mentions of water, soil and other environment parameters for "Areal" context and names and characteristics of diseases for "Pathogenicity" context

As it is followed from the table there is relatively small number of formal concepts in the contexts. This is due to the sparse form of all contexts generated by conceptual graphs.

For extracting relations we use visualization on the current stage of modeling technique. It allows getting results only for relatively small lattices.

Often relations between concepts in concept lattice may be treated as facts. Extracting facts from concept lattices is realized by forming special views constructed on the lattice and corresponded to certain property (intent in the lattice) or entity (extent in the lattice) on the set of bacteria. Every view is a sub lattice. It shows the links between concrete bacterium and its properties.

An example of such view as the fragment of lattice is shown on Fig. 4. The lattice on the Fig. 4 contains formal concepts related to the following bacteria: *Borrelia turicatae, Frankia, Legionella, Clamydophila, Thermoanaerobacter tengcongensis, Xanthomonas oryzae*. Highlighted view on the figure illustrates gram-negative property of bacteria. Such bacteria are resistant to conventional antibiotics.

Using this view, some facts about bacteria can be extracted:

- only three bacteria from the set, *Thermoanaerobacter tengcongensis, Clamydophila* and *Xanthomonas oryzae*, are gram-negative;
- two gram-negative bacteria, *Thermoanaerobacter tengcongensis* and *Xanthomonas oryzae*, have the shape as rod;
- one of gram-negative bacteria, *Clamydophila*, is obligately pathogenic.

Note that attribute *obligately pathogenic* was formed directly from the two words in the text according to the rule of marking words denoting extreme situation.

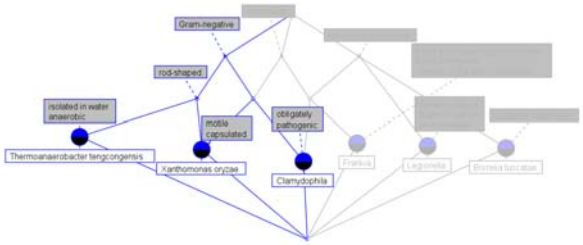


Figure 4 Example of view of gram-negative property of bacteria.

Comparing our results of relations extraction with the known ones from [23] we resume that concept lattice provides principally another variant of solution of this task. In [23] results of relations extraction are presented as marked words in the texts. Visualized concept lattice is more powerful object for investigating relations.

4 Conclusions and future work

This paper describes the idea of joining two paradigms of conceptual modeling - conceptual graphs and concept lattices. Current results of realizing this idea as CGs – FCA modeling on textual data show its good potential for knowledge extraction.

In spite of advantage of CGs – FCA modeling there are some problems which need to be solved for improving the quality of modeling technique.

1. Conceptual graphs acquired from texts contain many noise elements. Noise is constituted by the text elements that contain no useful information or cannot be interpreted as facts. Noise elements significantly decrease efficiency of algorithms of CGs – FCA modeling. To exclude noise we need to distinguish textual data that can be excluded from consideration, for example, information about when and by whom a bacterium was first identified.
2. Empirical rules which we use for creating formal contexts cannot embrace all configurations of conceptual graphs. More formal approach to creating formal contexts on the set of conceptual graphs will guarantee the completeness of solution. We guess that using patterns structures and their projections is that way of formalizing CGs – FCA modeling technique.
3. The next stage of developing CGs – FCA modeling is creating fledged information system which process user queries and produce solutions of certain tasks on textual data. Not only visualization but also special user oriented interfaces to concept lattice will be created in this system.

References

[1] BioNLP 2014. Workshop on Biomedical Natural Language Processing. Proceedings of the Workshop. The Association for Computational Linguistics. Baltimore, 2014. 155 p. http://acl2014.org/acl2014/W14-34/W14-34-2014.pdf

[2] Bogatyrev M., Latov, V., Stolbovskaya. Application of Conceptual Graphs in Digital Libraries. – Proc. RCDL (Digital libraries: advanced methods and technologies, digital collections), pp. 464-468. 2007.

[3] Bogatyrev M., Kolosoff A. Using Conceptual Graphs for Text Mining in Technical Support Services. Pattern Recognition and Machine Intelligence. - Lecture Notes in Computer Science, 2011, Volume 6744/2011, pp. 466-471. Springer-Verlag, Heidelberg, 2011. http://link.springer.com/chapter/10.1007%2F978-3-642-21786-9_75

[4] Bogatyrev, M., Nuriahmetov, V., Application of Conceptual Structures in Requirements Modeling. – Proc. of the International Workshop on Concept Discovery in Unstructured Data (CDUD 2011) at the Thirteenth International Conference on Rough Sets, Fuzzy Sets, Data Mining and Granular Computing - RSFDGrC 2011. Moscow, Russia, 2011, pp. 11-19.

[5] Bogatyrev, M. Y., Vakurin V. S. Conceptual Modeling in Biomedical Data Research. Mathematical Biology and Bioinformatics. 2013. Vol. 8. № 1, pp. 340–349. (in Russian). http://www.matbio.org/2013/Bogatyrev_8_340.pdf

[6] Bossy R, Jourde J, Manine A-P, Veber P, Alphonse E, Van De Guchte M, Bessières P, Nédellec C: BioNLP 2011 Shared Task - The Bacteria Track. BMC Bioinformatics. 2012, 13: S8, pp. 1-15. http://bmcbioinformatics.biomedcentral.com/articles/10.1186/1471-2105-13-S11-S3

[7] Campbell, N. A., etc., Biology: Concepts and Connections. Benjamin-Cummings Publishing Company, 2005.

[8] Carpineto, C., & Romano, G. Exploiting the Potential of Concept Lattices for Information Retrieval with CREDO. Journal of Universal Computing, 10, 8, 985-1013. 2004.

[9] Carpineto, C., Romano, G. Using Concept Lattices for Text Retrieval and Mining. In B. Ganter, G.

Stumme, and R. Wille (Eds.), Formal Concept Analysis: Foundations and Applications. Lecture Notes in Computer Science 3626, pp. 161-179. Springer-Verlag, Berlin, 2005.

[10] Cimiano, P. Hotho, A. Staab, S. Learning Concept Hierarchies from Text Corpora using Formal Concept Analysis. Journal of Artificial Intelligence Research, Volume 24, pp. 305-339. 2005. http://arxiv.org/pdf/1109.2140.pdf

[11] Edhlund, B., McDougall, A., Pubmed Essentials, Mastering the World's Health Research Database. Form & Kunskap AB, 2014. https://www.amazon.com/PubMed-Essentials-Mastering-Research-Database/dp/1312289457

[12] Ganter, B., Kuznetsov, S.O. Pattern structures and their projections. In: ICCS, pp. 129–142. 2001

[13] Ganter, B., Stumme, G., Wille, R., eds., Formal Concept Analysis: Foundations and Applications, Lecture Notes in Artificial Intelligence, No. 3626, Springer-Verlag. 2005

[14] Hunter L, Cohen KB. Biomedical language processing: what's beyond PubMed? Mol. Cell. 2006. V. 21. P. 589–594.

[15] Ignatov, Dmitry I. and Kuznetsov, Sergei O. and Poelmans, Jonas, Concept-based Biclustering for Internet Advertisement. In: Vreeken, J and Ling, C and Zaki, MJ and Siebes, A and Yu, JX and Goethals, B and Webb, G and Wu, X , Eds., Proc. of 12th IEEE International Conference On Data Mining Workshops (ICDMW 2012), pp. 123-130, 2012.

[16] Meštrović, A. Semantic Matching Using Concept Lattice. Concept Discovery in Unstructured Data, CDUD 2012, pp. 49-58. http://ceur-ws.org/Vol-871/paper_6.pdf

[17] Montes-y-Gomez, Gelbukh, Lopez-Lopez, Baeza-Yates, Flexible Comparison of Conceptual Graphs. Lecture Notes in Computer Science 2113. Springer-Verlag, 2001.

[18] Obitko, M., Snasel,V., Smid, Jan. Ontology Design with Formal Concept Analysis. – Proc. of the CLA 2004 International Workshop on Concept Lattices and their Applications. http://ceur-ws.org/Vol-110/paper12.pdf

[19] Olivé, Antoni, Conceptual Modeling of Information Systems. Springer-Verlag. Berlin, Heidelberg, 2007. https://www.amazon.com/dp/3540393897/ref=rdr_ext_tmb

[20] Otero P. G., Lopes G. P., Agustini, A., Automatic Acquisition of Formal Concepts from Text, Journal for Language Technology and Computational Linguistics. Vol. 23(1), pp. 59-74. 2008.

[21] Poelmans J., Kuznetsov S. O., Ignatov D. I., Dedene G. Formal Concept Analysis in knowledge processing: A survey on models and techniques // Expert Systems with Applications. 2013. Vol. 40. No. 16. P. 6601-6623.

[22] Priss, U., Linguistic Applications of Formal Concept Analysis. In: Ganter; Stumme; Wille (eds.), Formal Concept Analysis, Foundations and Applications. Springer Verlag. LNAI 3626, p. 149-160. 2005

[23] Ratkovic, Z., Golik, W., Warnier, P. Event extraction of bacteria biotopes: a knowledge-intensive NLP-based approach. - BMC Bioinformatics 2012, 13, (Suppl 11): S8, pp. 1-11. http://bmcbioinformatics.biomedcentral.com/articles/10.1186/1471-2105-13-S11-S8

[24] Sowa, J.F., Conceptual Structures: Information Processing in Mind and Machine. Addison-Wesley, London, UK. 1984

[25] U.S. National Library of Medicine. http://www.ncbi.nlm.nih.gov/pubmed .

[26] Wille, R. Conceptual Graphs and Formal Concept Analysis. Proceedings of the Fifth International Conference on Conceptual Structures: Fulfilling Peirce's Dream. 290 - 303. Springer-Verlag, London. 1997

[27] ConExp-NG. https://github.com/fcatools/conexp-ng

Методы анализа данных

Data analysis methods

Matrix Clustering Algorithms for Vertical Partitioning Problem: an Initial Performance Study

© Viacheslav Galaktionov[1]
viacheslav.galaktionov@gmail.com ,
© Boris Novikov[1]
b.novikov@spbu.ru ,
© George Chernishev[1,2]
g.chernyshev@spbu.ru,
© Dmitry A. Grigoriev[1]
d.a.grigoriev@spbu.ru

[1]Saint-Petersburg State University, Saint-Petersburg, Russia
[2]JetBrains Research, Saint-Petersburg, Russia

Abstract

Matrix clustering algorithms are among the oldest approaches to the vertical partitioning problem. They can be summarized as follows: (1) given a workload, construct an Attribute Usage Matrix (AUM), (2) apply some kind of a row and column permutation algorithm and (3) extract the resulting clusters which define the required fragments.
This naive approach holds some promise for a number of contemporary applications: (1) dynamization of vertical partitioning (2) big data applications and other cases of resource constraints (3) tuning of multistores.
In this paper we examine a number of existing matrix clustering algorithms used for vertical partitioning. We study these algorithms and assess the quality of the solutions. The experiments are run on the TPC-H workload using the PostgreSQL DBMS.

1 Introduction

The vertical partitioning problem [5] is one of the oldest problems in the database domain. There are dozens or even hundreds of studies available on the subject. It is a subproblem of the general database physical structure selection problem. It can be described as follows [9]: find a configuration (a set of vertical fragments) which would satisfy the given constraints and which will provide the best performance. There are two major classes of approaches to this problem:

- Cost-based approach [3, 16, 21, 33]. Studies that follow this approach construct a cost model, which is used to predict the performance of a workload for any given configuration. Next, an algorithm enumerating the configuration space is used.
- Procedural approach [29, 32, 35]. These studies do not use the notion of configuration cost. Instead, they propose some kind of a procedure which will result in a "good" configuration. Usually, these studies provide some intuitive explanation why the ensuing configuration would be "good".

The abundance of studies is justified by the following considerations:

- It was proved that the problem of vertical partitioning is an NP-Hard problem [4, 29, 37], just like many other physical design problems [6, 22, 37].
- Estimation errors related to both the system parameters and workload parameters. System parameters (hardware and software) in some cases cannot be measured precisely. Workload parameters can also be imprecise, e.g. not all queries are known in advance, or some of the known queries are not run. All these errors can cause the performance of the solution to deteriorate.

The procedural approach was very popular in the '80s and '90s because of the lack of computational resources. Nowadays, the interest for it has largely declined, and the majority of contemporary studies follows the cost-based one. This approach produces more accurate recommendations by incorporating additional information into the selection process. However, procedural approach has a number of promising applications:

- Dynamization of vertical partitioning [24, 28, 34, 36]. All of the previous vertical partitioning studies considered the problem in a static context, i.e. a configuration is selected once. In case of changes in the workload or the data the algorithm has to be re-run. In the new formulation the goal is to adapt the partitioning scheme to a constantly changing workload. The straightforward technique of the repeated re-run of a cost-based algorithm is not applicable due

Proceedings of the XVIII International Conference «Data Analytics and Management in Data Intensive Domains» (DAMDID/RCDL'2016), Ershovo, Russia, October 11 - 14, 2016

to its formidable costs of operation. Otherwise, its application will result in query processing stalls which should be avoided at all costs in this formulation. However, the procedural approach is not so computationally demanding as the cost-based one. Thus, low-quality solutions are acceptable as long as they provide improvement over the previous configuration and help us avoid queryprocessing stalls.

- Big data applications or any other cases featuring constrained resources.
- Tuning of multistores [27] or any other case when no details or only inaccurate estimates of physical parameters are known. It was already noted in the '80s [31] that the procedural approach is well-suited for such cases. A multistore system is a database system which consists of several distinct data stores, e.g. a Hadoop's HDFS and an RDBMS. This kind of a system is a modern example of the case where not every physical parameter of underlying data stores is known.

In this paper we evaluate a particular subclass of procedural vertical partitioning algorithms – the matrix clustering algorithms.

To the best of our knowledge, this study is the first one to evaluate this class of algorithms using a real DBMS and a real workload [1].

2 Related Work

2.1 Classification

The vertical partitioning problem is one of the oldest problems in the database domain. There are several dozens of studies on this topic, and most of them concern various algorithms. Several surveys can be found in the references [14, 15]. Vertical partitioning algorithms can be classified into two major groups: cost-based and procedural, where the latter employs three types of approaches:

- Attribute affinity and matrix clustering approaches [10, 11, 13, 19, 23]. In affinity-based approaches, closeness between every two attributes is first calculated, and then it is used to define the borders of the resulting fragments. This closeness is called attribute affinity. At the first step a workload is used to create an AUM, then an Attribute Affinity Matrix (AAM) is constructed using a paper-specific transformation procedure. Finally, a row and column permutation algorithm is applied. Matrix clustering approaches operate on the AUM and start with the permutation part.
- Graph approaches [12, 17. 29, 32, 39]. Most of the graph approaches treat the AAM as an adjacency matrix of an undirected weighted graph. In this graph nodes denote attributes and edges represent a bound's strength. Then a template is sought by various means, e.g. kruskal-like algorithms or hamiltonian way cut. The resulting templates are used to construct partitions.
- Data mining approaches [8, 20, 35]. This is a relatively new vertical partitioning technique that uses association rules to derive vertical fragments. Most of these works mine a workload (a transaction set) for rules which use sets of attributes as items. In these studies existing algorithms for association rule search are used to uncover relations between attributes. In particular, an adapted Apriori [2] algorithm is a very popular choice.

Let us review the matrix clustering approach in detail.

2.2 Matrix Clustering Approach

The general scheme of this approach is the following:

- Construct an Attribute Usage Matrix (AUM) from the workload. The matrix is constructed as follows:

$$M_{ij} = \begin{cases} 1, \text{ query } i \text{ uses attribute } j \\ 0, \text{ otherwise} \end{cases}$$

- Cluster the AUM by permuting its rows and columns to obtain a block diagonal matrix.
- Extract these blocks and use them to define the resulting partitions.

Some approaches do not operate on a 0-1 matrix. Instead they modify matrix values to account for additional information like query frequency, attribute size and so on.

Let us consider an example. Suppose that we have six queries accessing six attributes:

q1: SELECT a FROM T WHERE a > 10;
q2: SELECT b, f FROM T;
q3: SELECT a, c FROM T WHERE a = c;
q4: SELECT a FROM T WHERE a < 10;
q5: SELECT e FROM T;
q6: SELECT d, e FROM T WHERE d + e > 0;

The next step is the creation of an AUM using this workload. The resulting matrix is shown on Figure 1a. Having applied a matrix clustering algorithm, we acquire the reordered AUM (Figure 1b). The resulting fragments are the following: (a, b), (b, f), (d, e).

However, not all matrices are fully decomposable. Consider the matrix presented on Figure 2. The first column obstructs the perfect decomposition into several clusters. In this case, the algorithm should produce a decomposition which would minimally harm query processing and would result in an overall performance improvement. Matrix clustering algorithms employ different strategies to select such a decomposition.

	a	b	c	d	e	f
q1	1	0	0	0	0	0
q2	0	1	0	0	0	1
q3	1	0	1	0	0	0
q4	1	0	0	0	0	0
q5	0	0	0	0	1	0
q6	0	0	0	1	1	0

(a) AUM

	a	b	c	d	e	f
q1	1	0	0	0	0	0
q3	1	1	0	0	0	0
q4	1	0	0	0	0	0
q2	0	0	1	1	0	0
q6	0	0	0	0	1	1
q5	0	0	0	0	0	1

(b) Reordered AUM

Figure 1 Matrix clustering algorithm

2.3 Matrix Clustering Approach

The first study to introduce matrix clustering to vertical partitioning was the work of Hoffer [23]. The idea is to store together (in one file) attributes possessing identical retrieval patterns. The patterns are expressed through the notion of attribute cohesion, which shows how attributes in a pair are related to each other. The author proposes a pairwise attribute similarity measure to capture this cohesion.

The proposed measure relies on three parameters: co-access frequency of a pair of attributes, attribute length and relative importance of the query. This measure was designed having the following properties in mind: it is non-decreasing by co-access frequency, non-decreasing by both attribute lengths (individually) and the function is non-increasing in the combined length of attributes.

Finally, having an attribute affinity matrix, an existing clustering algorithm (Bond Energy Algorithm, BEA) [30] is used. It permutes rows and columns to maximize nearest neighbour bond strengths. The author was motivated in his choice by the following: this algorithm is insensitive to the order in which items are presented, it has a low computation time, etc. However, this algorithm has a disadvantage: it requires human attention for cluster selection.

a	b	c	d	e	f
1	1	1	0	0	0
1	1	1	0	0	0
1	1	1	0	0	0
1	0	0	1	1	0
1	0	0	1	1	0
1	0	0	0	0	1

Figure 2 Non-decomposable matrix

BEA is not the only existing matrix clustering algorithm. Another permutation algorithm was proposed in the reference [38]. Similarly to BEA, it permutes rows and columns, but tries to minimize the spanning path of the graph represented by the original matrix. The improvement of these two algorithms is presented in the reference [7]. This algorithm is called the matching algorithm and it uses Hamming distance to produce clusters. According to the reference [10], the study [25] presents the Rank Order algorithm. Its idea is to sort rows and columns of the original matrix in descending order of their binary weight. The Cluster Identification (CI) algorithm by Kusiak and Chow [26] is an algorithm for clustering 0-1 matrices. The proposed approach is to detect clusters one by one using a special procedure. This procedure resembles the search of a transitive closure for rows and columns. It is an optimal algorithm that can solve the problem when the matrix is perfectly separable, e.g. when clusters do not intersect (there is no attribute sharing).

All of the aforementioned algorithms (except BEA) are generic matrix clustering algorithms. They do not address the vertical partitioning problem and do not even bear any database specifics. The next studies by Chun-Hung Cheng [10, 11, 13] attempt to apply matrix clustering approach to the database domain. Several new vertical partitioning algorithms were developed in his works. Let us consider them.

Chun-Hung Cheng criticizes existing matrix clustering algorithms [10, 11]:

- They do not always produce a solution matrix in a diagonal submatrix structure. Thus, these algorithms may require additional computation to extract them;
- These algorithms may require decision of database administrator to identify inter-submatrix attributes [10].

The first study [10] extends the original CI [26] algorithm to non-decomposable matrices. The proposed approach is to remove columns obstructing the decomposition (inter-submatrix attributes).

The author considered the following problem formulation *P*1 [10]: remove columns to decompose a matrix into separable submatrices with the maximum number of "1" entries retained in submatrices subject to the following constraints:

- C1: A submatrix must contain at least one row;
- C2: The number of rows in a submatrix cannot exceed upper limit, b;
- C3: A submatrix must contain at least one column.

In order to solve the problem, the branch and bound approach was used. This approach uses an objective function which maximizes the number of "1" entries in the resulting submatrices. During the tree traversal, upper and lower bounds are calculated and used to guide the enumeration process.

However, the basic approach required traversal of too many nodes, so the author augmented it with the following heuristic. A so-called **blocking measure** is calculated for each column. It estimates the likelihood of a column being an obstacle to the further decomposition of the matrix. Basically, it is the number of columns that would be involved in all queries which use the given

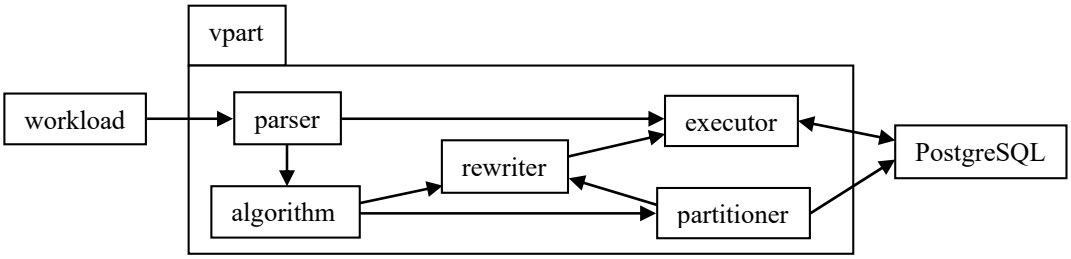


Figure 3 The architecture of our approach

attribute. Next, the columns are ordered by their respective values and the ones with the highest values are checked.

The study [11] also extends the original CI algorithm. The author adopts the same branch and bound approach as in his previous paper [10]. However, instead of the **blocking measure** a new **void measure** is developed. It has the same purpose, which is the estimation of the likelihood of a column being an inter-submatrix column. Essentially, this measure is the calculated "free space" to the left and to the right of the candidate cluster.

The next study of the author [13] addresses several shortcomings of his previous works:

- The problem of the parameter b. While this parameter helps prevent the formation of the huge clusters, it does not guarantee any quality of the resulting clusters. Also, the problem will have to be reformulated if several clusters of different sizes are needed.
- The dangling transaction problem. Applying the previous algorithm [11] a transaction not belonging to any cluster may be acquired: all of its attributes would be removed. Two examples are presented in the original paper.
- The previous work did not include such an important parameter as the access frequency of the transactions.

Thus, a new formulation *P*3 is proposed [13]: remove a minimal number of "1" entries to decompose a transaction-attribute access matrix into separable submatrices subject to the following constraints:

- C7: Transactions with all "0" entries in a submatrix are not allowed.
- C8: Attributes with all "0" entries in a submatrix are not allowed.
- C9: The **cohesion measure** of a submatrix is more than or equal to a threshold, δ.

Cohesion measure of a submatrix is the ratio of "1" elements to "0" elements. This new measure is used to ensure the quality of a cluster.

The problem is also solved with the branch and bound approach, again, the **void measure** is used to guide the order of node traversal.

Furthermore, in this work the author shows why dangling transactions should be avoided: an example is provided showing a case where it is possible to lose information regarding a cluster. Finally, the author extended his CI framework to consider query frequencies. This *P*4 formulation is the same as *P*3, but features a weighted sum of accesses [13]: minimize the loss of total accesses ($\sum_i \sum_j a_{ij} \times freq_i$) due to the removal of a_{ij} for decomposing a transaction-attribute matrix into separable submatrices subject to the same constraints C7–C9.

In this paper we study the approaches described in the references [10, 11, 13].

3 System Architecture

We have developed a program for experimental evaluation of the considered algorithms. Its architecture is presented on Figure 3. It consists of the following modules:

- The parser reads the workload from a file. It extracts the queries and passes them to the executor, so that their execution times can be measured. It also constructs the AUM, which serves as input for the selected algorithm.
- The algorithm identifies clusters and passes that information to the partitioner to create corresponding temporary tables.
- The query rewriter also receives this information. It replaces the name of the original table with the ones that were generated by the partitioner. It can handle subqueries; view support is not implemented yet.
- The partitioner generates new names and sends partitioning commands to the database. The exact commands are SELECT INTO and ALTER TABLE. The latter lets it transfer primary keys.
- The executor accepts queries and sends them to PostgreSQL to measure the time of execution.

4 Parallelization

Having implemented this system, we noticed unacceptable run times even for relatively small matrices. The author of these algorithms states that this is not a problem because the algorithm finds a good solution quickly and spends the major portion of its time just by checking the rest of the tree.

However, we decided to parallelize all of the algorithms. We managed to achieve this in a generic fashion, i.e. we applied a generic parallelization scheme for all of the branch and bound algorithms. In order to

Type	Q1	Q6	Q14	Q19
Original	11694	1365	1412	1663
partitioned	31558	1602	1379	1797

Figure 4 Scenario 1 – A09, QS1, 0.7

Type	Q6	Q14	Q19
Original	1439	1405	1673
Partitioned	1343	1093	2731

Figure 6 Scenario 2 – A09, QS2, 0.7

	1	2	3	4	5	6	7	8	9	10
Q1	0	1	1	1	1	1	1	1	0	0
Q6	0	1	1	1	0	0	0	1	0	0
Q14	1	0	1	1	0	0	0	1	0	0
Q19	1	1	1	1	0	0	0	0	1	1

(a) Original

	1	3	9	10	2	4	5	6	7	8
Q1	0	*	0	0	1	1	1	1	1	1
Q6	0	*	0	0	1	1	0	0	0	1
Q14	1	1	0	0	0	*	0	0	0	*
Q19	1	1	1	1	*	*	0	0	0	0

(b) Result

Figure 5 Scenario 1 – A09, QS1, 0.7

	1	2	3	4	5	6	7
Q6	0	1	1	1	1	0	0
Q14	1	0	1	1	1	0	0
Q19	1	1	1	1	0	1	1

(a) Original

	1	2	3	4	5	6	7
Q6	0	1	1	1	1	0	0
Q14	1	0	1	1	1	0	0
Q19	*	*	*	*	0	1	1

(b) Result

Figure 7 Scenario 2 – A09, QS2, 0.7

type	Q6	Q14
original	1555	1395
partitioned	1421	1062

Figure 8 Scenario 3 – A09, QS3, 0.7

type	Q6	Q14	Q19
original	1438	1360	1685
partitioned	1405	1148	1704

Figure 9 Scenario 4 – A09, QS2, 0.5

type	Q6	Q14	Q19
original	1385	1409	1648
partitioned	2201	1377	1855

Figure 10 Scenario 5 – A09, QS2, 0.9

Type	Q1	Q6	Q14	Q19
original	11515		1367	1608
partitioned	31173	934	1479	1989

Figure 12 Scenario 6 – A95, QS1, 2

implement it we employed the Threading Building Blocks (TBB)[1].

Using these primitives, our already existing sequential implementation was parallelized with minimal effort. We replaced the explicit stack used in sequential depth-first traversal with TBB constructs[2]. Thus, we kept the node inspection code unchanged.

For the detailed information regarding the parallelization method and the results see the original paper [18].

5 Experiments

We have implemented three recent matrix clustering algorithms [10, 11, 13] (A94, A95, A09) and used PostgreSQL DBMS to evaluate them. Our experiments were conducted using the standard benchmark – TPC-H with scale factor 1. We measured the run times for original and partitioned configurations.

5.1 Hardware and Software Setups

In our experiments the following setup was used:

- PostgreSQL 9.5.2,
- Gentoo Linux (kernel 4.1.12),
- Intel® Core™ i7-3630QM (4 physical cores, hyper-threading enabled)
- 8GB (DDR3) RAM,
- GCC 4.9.3.

The database was placed in the main memory of the machine. In order to accomplish this, the PostgreSQL data directory was put on tmpfs, created with standard GNU/Linux utilities.

To ensure the reproducibility of our results we used sequential versions of algorithms for all comparisons. All of the workloads were executed sequentially.

5.2 Data Setup

For our evaluation we have chosen the LINEITEM and PART tables. Based on these tables we have formulated the following query setups:

- Query Setup 1 (QS1): Q1, Q6, Q14, Q19;
- Query Setup 2 (QS2): Q6, Q14, Q19;
- Query Setup 3 (QS3): Q6, Q14.

This is the initial series of experiments, so we tried to use simple scenarios. In these experiments we assume uniform distribution of query frequencies.

The author of the studied algorithms indicated that there are three possible strategies for dealing with inter-submatrix attributes: forming a separate cluster for all inter-submatrix attributes, duplicating them to every subrelation and keeping them in the relation which uses them more often. He argues that the decision which

[1] https://www.threadingbuildingblocks.org/

[2] https://www.threadingbuildingblocks.org/docs/help/reference/task_scheduler.htm

strategy to apply is usually left to database administrator. In this study we employ the first strategy.

5.3 Scenario 1

In this experiment we used the most recent algorithm from the reference [13] (A09). The cohesion parameter

	1	2	3	4	5	6	7
Q6	0	1	1	1	1	0	0
Q14	1	0	1	1	1	0	0
Q19	1	1	1	1	0	1	1

(a) Original

Figure 11 Scenario 5 – A09, QS2, 0.9

	1	2	3	4	5	6	7	8	9	10
Q1	0	1	1	1	1	1	1	1	0	0
Q6	0	1	1	1	0	0	0	1	0	0
Q14	1	0	1	1	0	0	0	1	0	0
Q19	1	1	1	1	0	0	0	0	1	1

(a) Original

Figure 13 Scenario 6 – A95, QS1, 2

was set to 0.7 and QS1 was used. Table 4 contains the performance for this scenario.

As we can see, only run time for Q14 improved and the overall time significantly increased. The query Q1 can be characterized by a large number of aggregates and read attributes. This is the possible reason for such performance deterioration. The partitioning scheme is presented in Table 5.

5.4 Scenario 2

This experiment also addresses the A09 algorithm with the same cohesion parameter. However, we decided to discard Q1 from the workload to check whether that would improve the overall performance. The results are presented in Table 6. While Q6 and Q14 performance improved, the Q19 performance has greatly deteriorated. The net gain is also negative in this case. The corresponding partitioning scheme is presented in Table 7.

5.5 Scenario 3

In this scenario we examined A09 on the QS3. The results are shown in Table 8. We do not demonstrate the original and partitioned matrices due to the space constraints and due to the fact that the algorithm returned only one cluster, which was identical to the input one. Thus, overall improvement was achieved via transfer of all of untouched attributes to a separate cluster.

5.6 Scenario 4

In this experiment we again considered A09 on QS2, but lowered the cohesion value to 0.5. Similarly, we obtained a positive net gain (see Table 9). Unfortunately, input and output matrices indicate that the reason for this improvement is the same as in Scenario 3.

5.7 Scenario 5

Here we evaluate the behavior of A09 with QS2 and cohesion value of 0.9. Results are presented in Tables 10 and 11. There is also negative overall gain.

5.8 Scenario 6

	1	2	6	7	3	4	5
Q6	0	*	0	0	1	1	1
Q14	*	0	0	0	1	1	1
Q19	1	1	1	1	*	*	0

(b) Result

	1	9	10	5	6	7	2	3	4	8
Q1	0	0	0	1	1	1	1	1	1	1
Q6	0	0	0	0	0	0	1	1	1	1
Q14	1	0	0	0	0	0	0	1	1	1
Q19	1	1	1	0	0	0	1	1	1	0

(b) Result

In this experiment we tried a different algorithm – A95 on QS1 with the maximum number of rows being 2. The outcome is presented in Tables 12, 13.

5.9 Other Scenarios and Results

We have also conducted a number of other experiments, but unfortunately, we are limited by the space available. Here is a brief summary of our findings:

- If we select a lot of attributes in one query of the workload, these algorithms will perform poorly;
- These algorithms perform well on workloads which have several columns consisting of "0" entirely (containing no accesses in the workload);
- It may be beneficial to set a low cohesion value in order to achieve better performance. This is accomplished by eliminating additional joins;
- Algorithms A95, A94 and Optimal exhibit the similar behavior during our tests;
- There are cases when any of the algorithms (Optimal, A94, A95) can return no solution;
- In order to obtain a non-trivial solution cohesion parameter should be higher than one of the original matrix.

6 Conclusions

In this paper we have studied three newer matrix clustering algorithms [10, 11, 13]. We have implemented these algorithms and used PostgreSQL with TPC-H workload to evaluate them. In our experiments we employed one of several inter-cluster attribute handling strategies. Preliminary results suggest that all of these algorithms perform poorly in this environment, often yielding partitioning schemes worse than the original one.

References

[1] TPC Benchmark H. Decision Support. Version 2.17.1. http://www.tpc.org/tpch [Accessed: 2016 01 08]

[2] R. Agrawal and R. Srikant. Fast algorithms for mining association rules in large databases. In Proceedings of the 20th International Conference on Very Large Data Bases, VLDB '94, pages 487–499, San Francisco, CA, USA, 1994. Morgan Kaufmann Publishers Inc.

[3] S. Agrawal, V. Narasayya, and B. Yang. Integrating vertical and horizontal partitioning into automated physical database design. In Proceedings of the 2004 ACM SIGMOD international conference on Management of data, SIGMOD '04, pages 359–370, New York, NY, USA, 2004. ACM.

[4] P. M. G. Apers. Data allocation in distributed database systems. ACM Trans. Database Syst., 13:263–304, 1988.

[5] L. Bellatreche. Optimization and tuning in data warehouses. In L. LIU and M. ÖZSU, editors, Encyclopedia of Database Systems, pages 1995–2003. Springer US, 2009.

[6] L. Bellatreche, K. Boukhalfa, and P. Richard. Data partitioning in data warehouses: Hardness study, heuristics and oracle validation. In I.-Y. Song, J. Eder, and T. Nguyen, editors, Data Warehousing and Knowledge Discovery, volume 5182 of Lecture Notes in Computer Science, pages 87–96. Springer Berlin / Heidelberg, 2008. 10.1007/978-3-540-85836-2_9.

[7] M. V. Bhat and A. Haupt. An efficient clustering algorithm. Systems, Man and Cybernetics, IEEE Transactions on, SMC-6(1):61–64, 1976.

[8] M. Bouakkaz, Y. Ouinten, and B. Ziani. Vertical fragmentation of data warehouses using the FP-Max algorithm. In Innovations in Information Technology (IIT), 2012 International Conference on, pages 273–276, march 2012.

[9] S. Chaudhuri and G. Weikum. Self-management technology in databases. In L. Liu and M. Özsu, editors, Encyclopedia of Database Systems, pages 2550–2555. Springer US, 2009.

[10] C. Cheng. Algorithms for vertical partitioning in database physical design. Omega, 22(3):291–303, 1994.

[11] C.-H. Cheng. A branch and bound clustering algorithm. Systems, Man and Cybernetics, IEEE Transactions on, 25(5):895–898, 1995.

[12] C.-H. Cheng, W.-K. Lee, and K.-F. Wong. A genetic algorithm-based clustering approach for database partitioning. Systems, Man, and Cybernetics, Part C: Applications and Reviews, IEEE Transactions on, 32(3):215–230, 2002.

[13] C.-H. Cheng and J. Motwani. An examination of cluster identification-based algorithms for vertical partitions. Int. J. Bus. Inf. Syst., 4(6):622–638, 2009.

[14] G. Chernishev. Towards self-management in a distributed column-store system. In T. Morzy, P. Valduriez, and L. Bellatreche, editors, New Trends in Databases and Information Systems, volume 539 of Communications in Computer and Information Science, pages 97–107. Springer International Publishing, 2015.

[15] G. Chernishev. Vertical Partitioning in Relational DBMS. Talk at the Moscow ACM SIGMOD chapter meeting; slides and video: http://synthesis.ipi.ac.ru/sigmod/seminar/s20150430 30 4 2015 [Accessed: 2016 01 08].

[16] W. Chu and I. Ieong. A transaction-based approach to vertical partitioning for relational database systems. Software Engineering, IEEE Transactions on, 19(8):804–812, 1993.

[17] J. Du, K. Barker, and R. Alhajj. Attraction — a global affinity measure for database vertical partitioning. In ICWI, pages 538–548. IADIS, 2003.

[18] V. Galaktionov. Parallelization of matrix clustering algorithms (accepted). In Proceedings of the Sixth International Conference on Informatics Problems (SPISOK 2016), 2016.

[19] N. Gorla and W. J. Boe. Database operating efficiency in fragmented databases in mainframe, mini, and micro system environments. Data & Knowledge Engineering, 5(1):1–19, 1990.

[20] N. Gorla and B. P. W. Yan. Vertical fragmentation in databases using data-mining technique. In J. Erickson, editor, Database Technologies: Concepts, Methodologies, Tools, and Applications, pages 2543–2563. IGI Global, 2009.

[21] M. Hammer and B. Niamir. A heuristic approach to attribute partitioning. In Proceedings of the 1979 ACM SIGMOD international conference on Management of data, SIGMOD '79, pages 93–101, New York, NY, USA, 1979. ACM.

[22] V. Harinarayan, A. Rajaraman, and J. D. Ullman. Implementing data cubes efficiently. In Proceedings of the 1996 ACM SIGMOD International Conference on Management of Data, SIGMOD'96, pages 205–216, New York, NY, USA, 1996. ACM.

[23] J. A. Hoffer and D. G. Severance. The use of cluster analysis in physical data base design. In Proceedings of the 1st International Conference on Very Large Data Bases, VLDB '75, pages 69–86, New York, NY, USA, 1975. ACM.

[24] Jindal and J. Dittrich. Relax and let the database do the partitioning online. In M. Castellanos, U. Dayal, and W. Lehner, editors, Enabling Real-Time Business Intelligence, volume 126 of Lecture Notes in Business Information Processing, pages 65–80. Springer Berlin Heidelberg, 2012.

[25] J.R. King Machine-component grouping in production flow analysis: an approach using a rank order clustering algorithm. Int. J. Prod. Res., 18(2):213–232, 1980.

[26] Kusiak and W. Chow. An efficient cluster identification algorithm. Systems, Man and

Cybernetics, IEEE Transactions on, SMC-17(4):696–699, 1987.

[27] J. LeFevre, J. Sankaranarayanan, H. Hacigumus, J. Tatemura, N. Polyzotis, and M. J. Carey. MISO: Souping up big data query processing with a multistore system. In Proceedings of the 2014 ACM SIGMOD International Conference on Management of Data, SIGMOD '14, pages 1591–1602, New York, NY, USA, 2014. ACM.

[28] L. Li and L. Gruenwald. Self-managing online partitioner for databases (SMOPD): A vertical database partitioning system with a fully automatic online approach. In Proceedings of the 17th International Database Engineering & Applications Symposium, IDEAS '13, pages 168–173, New York, NY, USA, 2013. ACM.

[29] X. Lin, M. Orlowska, and Y. Zhang. A graph based cluster approach for vertical partitioning in database design. Data & Knowledge Engineering, 11(2):151–169, 1993.

[30] W. McCormick, P. Schweitzer, and W. White. Problem decomposition and data reorganization by a clustering technique. Oper. Res., 20(5):993–1009, 1972.

[31] S. Navathe, S. Ceri, G. Wiederhold, and J. Dou. Vertical partitioning algorithms for database design. ACM Trans. Database Syst., 9:680–710, 1984.

[32] S. B. Navathe and M. Ra. Vertical partitioning for database design: a graphical algorithm. In Proceedings of the 1989 ACM SIGMOD international conference on Management of data, SIGMOD '89, pages 440–450, New York, NY, USA, 1989. ACM.

[33] S. Papadomanolakis and A. Ailamaki. An integer linear programming approach to database design. In Proceedings of the 2007 IEEE 23rd International Conference on Data Engineering Workshop, ICDEW '07, pages 442–449, Washington, DC, USA, 2007. IEEE Computer Society.

[34] L. Rodrìguez and X. Li. A dynamic vertical partitioning approach for distributed database system. In Systems, Man, and Cybernetics (SMC), 2011 IEEE International Conference on, pages 1853–1858, 2011.

[35] L. Rodrìguez and X. Li. A support-based vertical partitioning method for database design. In Electrical Engineering Computing Science and Automatic Control (CCE), 2011 8th International Conference on, pages 1–6, oct. 2011.

[36] L. Rodrìguez, X. Li, and P. Mejìa-Alvarez. An active system for dynamic vertical partitioning of relational databases. In I. Batyrshin and G. Sidorov, editors, Advances in Soft Computing, volume 7095 of Lecture Notes in Computer Science, pages 273–284. Springer Berlin Heidelberg, 2011.

[37] D. Sacca and G. Wiederhold. Database partitioning in a cluster of processors. ACM Trans. Database Syst., 10:29–56, 1985.

[38] J. R. Slagle, C. L. Chang, and S. R. Heller. A clustering and data-reorganizing algorithm. Systems, Man and Cybernetics, IEEE Transactions on, SMC-5(1):125–128, Jan 1975.

[39] J. H. Son and M.-H. Kim. α-partitioning algorithm: Vertical partitioning based on the fuzzy graph. In Proceedings of the 12th International Conference on Database and Expert Systems Applications, DEXA '01, pages 537–546, London, UK, UK, 2001. Springer-Verlag.

Разработка ансамбля алгоритмов кластеризации на основе изменяющихся метрик расстояний

© П. В. Бочкарёв © В. С. Киреев

Национальный исследовательский ядерный университет «МИФИ»,
Москва, Российская Федерация
pvbochkarev@mephi.ru vskireev@mephi.ru

Аннотация

В настоящее время происходит активное накопление данных большого объёма в различных информационных средах, таких как социальные, корпоративные, научные и другие. Интенсивное использование больших данных в различных областях стимулирует повышенный интерес исследователей к развитию методов и средств обработки и анализа массивных данных огромных объёмов и значительного многообразия. Одним из перспективных направлений в аналитике интенсивных данных является кластерный анализ, который позволяет решить такие задачи как, сокращение размерности исходного набора данных, выявление паттернов и т.д. В данной статье авторами предлагается ансамбль алгоритмов кластеризации, состоящий из базовых алгоритмов K-means, отличающихся по одному параметру - метрике расстояния между объектами. Для оценки работы разработанного ансамбля использованы открытые данные архива UCI.

1 Введение

Данные большого объёма (BigData), используются в различных процессах, таких как извлечение информации из веб-ресурсов, выявления общих закономерностей в областях с интенсивным использованием данных и т.д. Эти данные необходимо структурировать, классифицировать, подвергать тщательному анализу. В этом случае кластерный анализ является основой многих научных исследований [14]. Кластеризация (от англ. cluster – скопление), это сегментация через выделение определённых объединений однородных элементов, которые рассматриваются как самостоятельные единицы, обладающие определёнными свойствами [15].

В результате процедуры кластеризации образуются «кластеры», то есть группы очень похожих объектов [16].

Труды XVIII Международной конференции DAMDID/RCDL'2016 «Аналитика и управление данными в областях с интенсивным использованием данных», Ершово, 11-14 октября 2016

Под критерием качества кластеризации обычно понимается некоторый функционал, зависящий от разброса объектов внутри группы и расстояний между ними [9].

Кластеризация отличается от классификации тем, что изначально неизвестны ни количество, ни свойства классов (кластеров). К особенностям кластеризации можно отнести следующее:

- возможность определения заранее неизвестного класса объектов по начальным характеристикам;
- возможность обработки сколь угодно большого количества объектов в достаточно короткие сроки.

Устойчивость решений в задачах кластеризации может быть повышена благодаря формированию ансамбля алгоритмов [13] и построению с его помощью коллективного решения на основе мнений участников ансамбля, где под мнением алгоритма подразумевается его вариант разбиения данных на кластеры.

Данные свойства кластерного анализа особо актуальны при работе в областях с интенсивным использованием данных, когда предметная область слабо формализована, например, для анализа текстовых документов, изображений и т.д.

Основное внимание в данной работе уделяется построению ансамбля алгоритмов кластеризации на основе изменяющихся метрик расстояний для анализа данных большого объема.

2 Современные подходы к решению проблемы

Выбор метода кластеризации зависит от количества данных и от того, требуется ли обрабатывать и анализировать несколько типов данных одновременно [6, 10].

На практике чаще всего используются гибридные подходы, в которых шлифование кластеров выполняется методом K-средних (см. форм.1), а начальное разбиение – одним из более универсальных и мощных методов.

$$V = \sum_i^k \sum_{x_j \epsilon S_i} \left(x_j - \mu_i\right)^2, \qquad (1)$$

где k – число кластеров, S_i – полученные кластеры, $i=1, 2,\ldots, k$ и μ_i – центры масс векторов.

Данные о сравнении алгоритмов представлены в таб.1 [7].

Таблица 1 Сравнительная таблица алгоритмов

Алгоритм кластеризации	**Входные данные**	**Результаты**
иерархическая	число кластеров или порог расстояния для усечения иерархии	бинарное древо кластеров
K-средних	число кластеров	центры кластеров
C-средних	число кластеров, степень нечеткости	центры кластеров, матрица принадлежности
выделение связных компонент	порог расстояния R	древовидная структура кластеров

Для определения расстояния между объектами в кластерном анализе используются различные метрики расстояний между объектами *x* и *x'*. Наиболее востребованными в кластерном анализе являются следующие метрики:

1. Евклидово расстояние
$$p(x,x') = \sqrt{\sum_i^n (x_i - x_i')^2}; \quad (2)$$
2. Манхэттенское расстояние
$$p(x,x') = \sum_i^n |x_i - x_i'|; \quad (3)$$
3. Расстояние Чебышева
$$p(x,x') = max(|x_i - x_i'|); \quad (4)$$
4. Коэффициент Жаккара
$$K(x,x') = \frac{\sum_i^n x_i x_i'}{\sum_i^n x_i^2 + \sum_i^n x_i'^2 - \sum_i^n x_i x_i'}; \quad (5)$$
5. Динамическая трансформация временной шкалы (dynamic time wrapping, DTW)
$$DTW(x,x') = \frac{\min\{\sum_{k=1}^K d(\omega_k)\}}{K}, \quad (6)$$
где *K* – длина пути между *x* и *x'*, который вычисляется по специальной матрице трансформаций [11].

Выбор метрики существенно влияет на качество кластеризации.

В настоящее время в кластерном анализе проявляется тенденция к применению коллективных методов [1]. Ранее было отмечено, что алгоритмы кластерного анализа не являются универсальными: каждый алгоритм имеет свою особую область применения (таблица 1). В том случае, если рассматриваемая область содержит различные типы данных, для выделения кластеров необходимо применять не один определённый алгоритм, а набор различных алгоритмов.

Ансамблевый (коллективный) подход позволяет снизить зависимость конечного решения от выбранных параметров исходных алгоритмов и получить более устойчивое решение даже при большом количестве шумов и выбросов в данных [9].

Существуют следующие основные методики получения ансамбля алгоритмов (см. Рис. 1)[8]:

1. нахождение консенсусного разбиения, т.е. согласованного разбиения при имеющихся нескольких решениях, оптимального по некоторому критерию;
2. вычисление согласованной матрицы сходства/различий (co-occurrence matrix).

Рисунок 1 Ансамбли алгоритмов кластеризации

При формировании окончательного решения используются результаты, полученные различными алгоритмами, либо одним алгоритмом с различными значениями параметров, по разным подсистемам переменных и т.д. В настоящее время ансамблевый подход является одним из наиболее перспективных направлений в кластерном анализе.

Примерами использования ансамбля алгоритмов кластеризации могут служить следующие. Созданный на основе непараметрического алгоритма MeanSC, ансамбль позволил улучшить показатели кластеризации многоканальных изображений [13]. А также, используя ансамбль алгоритмов кластеризации на основе K-средних и алгоритма SVM (Support Vector Machines), удалось повысить точность обнаружения сердечных аномалий, что позволило сократить время установления диагноза [2].

Таким образом, применяя ансамбль с различными наборами алгоритмов, в соответствии с их преимуществами и особенностями, можно создать наиболее подходящую схему кластеризации для определённой предметной области. Ранее также указывалось, что важным фактором, влияющим на результат кластеризации, является выбор конкретной метрики расстояний между объектами. Объединяя эти два подхода, можно существенно повысить эффективность кластерного анализа.

3 Предлагаемый подход

3.1 Ансамбль алгоритмов кластеризации

Предлагаемый авторами ансамбль алгоритмов представляет собой сочетание последовательных алгоритмов K-средних, каждый из которых предлагает свое разбиение, и иерархического агломеративного алгоритма, объединяющего

полученные решения с помощью особого механизма. В отличие от ансамбля, использующего алгоритм MeansSC [13], предложенный ансамбль опирается на результаты предварительного исследования исходных данных, которые представляют собой небольшой набор размеченных экспертами объектов. Минимально необходимый процент объема исходной выборки, гарантирующий заданную точность, подлежит дальнейшему изучению. Для определенности, в данной работе используется 0.5%, что в случае увеличения объема данных, очевидно, должно подлежать пересмотру.

На первом шаге каждый алгоритм K-средних, разбивает данные на кластеры, используя свою метрику расстояния. Затем, рассчитывается точность и вес мнения алгоритма в ансамбле по формуле 7:

$$\omega_l = \frac{Acc_l}{\sum_{l=1}^{L} Acc_l}, \quad (7)$$

где Acc_l – точность алгоритма l, т.е. отношение количество правильно кластеризованных объектов к объему всей выборки, а L – количество алгоритмов в ансамбле.

Для каждого полученного разбиения составляется предварительная бинарная матрица различий размера n x n, где n-количество объектов, необходимая для определения, занесены ли объекты разбиения в один класс. Затем рассчитывается согласованная матрица различий, каждый элемент которой представляет собой взвешенную (с использованием веса из формулы 7) сумму элементов предварительных матриц. Полученная матрица используется в качестве входных данных для алгоритма иерархической агломеративной кластеризации. Затем с помощью обычных приемов, таких как определение скачка расстояния агломерации, можно выбрать наиболее подходящее кластерное решение. Процедура создания ансамбля алгоритмов представлена на рисунке 2.

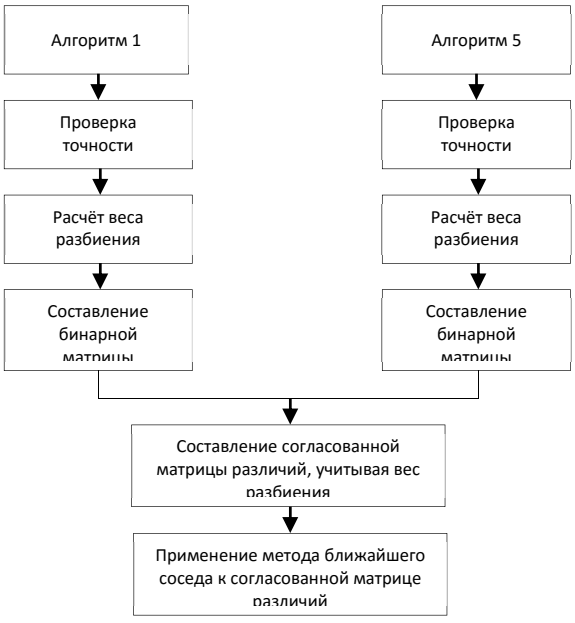


Рисунок 2 Ансамбль алгоритмов кластеризации

3.2 Алгоритмы кластеризации

В данном ансамбле алгоритмов кластеризации были использованы пять K-средних (см. форм.1), как один из наиболее востребованных алгоритмов кластеризации больших данных [12]. Для данных алгоритмов были использованы такие метрики, как:

- Евклидово расстояние (см. форм. 2)
- Манхэттенское расстояние (см. форм. 3)
- Расстояние Чебышева (см. форм. 4)
- Коэффициент Жаккара (см. форм. 5)
- DTW расстояние (см. форм. 6)

3.3 Наилучшее разбиение на кластеры

Для получения наилучшего разбиения на кластеры необходимо, как было упомянуто выше, составить бинарную матрицу сходства\различий на каждое L разбиение в ансамбле:

$$H_i = \{h_i(i,j)\}, \quad (8)$$

где $h_i(i, j)$ равен нулю, если элемент i и элемент j попали в один кластер, и 1 если нет.

Следующим шагом в составлении ансамбля алгоритмов кластеризации является составление согласованной матрицы бинарных разбиений.

$$H^* = \{h^*(i,j)\}, \quad (9)$$

$$h^*(i,j) = \sum_{l=1}^{L} w_l h_l(i,j), \quad (10)$$

где w_l – вес алгоритма.

Для формирования наилучшего разбиения по согласованной матрице был выбран алгоритм ближайшего соседа.

4 Валидация предлагаемого ансамбля

Для тестирования и оценки ансамбля алгоритмов кластеризации использовалось программное средство RapidMiner [3]. С помощью RapidMiner можно решать, как исследовательские (модельные), так и прикладные (реальные) задачи интеллектуального анализа данных, включая анализ текста, анализ мультимедиа, анализ потоков данных, что подходит для тестирования ансамбля алгоритмов кластеризации. В качестве данных для кластеризации использовались открытые данные с web-сайта UCI [4]. Данный пример содержит информацию о платежах клиентов с помощью пластиковых карт, всего 30 тыс. записей и 24 атрибута. Данные были размечены экспертным способом на 2 кластера, и эти результаты были взяты в качестве корректного решения.

Ниже представлены элементы схемы эксперимента, разработанной в RapidMiner. Для снижения размерности исходных данных был выбран метод главных компонентов (Principal component analysis, PCA) (см. Рис. 3). В качестве критерия выбора количества компонент был выбран критерий Кайзера (собственное значение компоненты больше единицы).

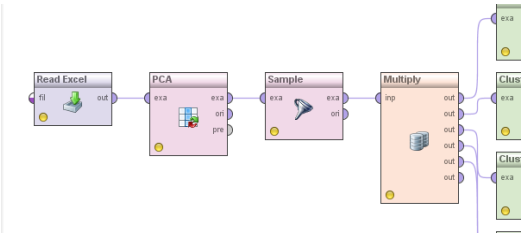


Рисунок 3 Снижение размерности данных

На следующем шаге была выявлена точность каждого алгоритма путём сравнения полученного разбиения на два кластера каждым алгоритмом с кластерами, размеченными экспертным способом. После получения значения точности каждого алгоритма, по формуле 7 был рассчитан вес мнения алгоритма (см. Рис. 4). Так, из графика видно, что наибольшим весом обладает алгоритм, использовавший расстояние Чебышева, а наименьшим весом - алгоритм с метрикой Жаккара.

Рисунок 4 Диаграмма веса алгоритмов

Далее была проведена кластеризация данных каждым алгоритмом. На следующем шаге, используя возможности RapidMiner, по формуле 8 были получены бинарные матрицы разбиения

На основе полученных результатов можно определить значение индекса качества группировки (вес разбиения). Используя вес каждого разбиения и сумму значений бинарных матриц сходства\различий (см. форм. 9), была составлена согласованная матрица различий для ансамбля алгоритмов кластеризации (см. форм. 10).

Применяя алгоритм ближайшего соседа к рассчитанной матрице, с помощью возможностей Rapidminer, было определено наилучшее разбиение.

На рисунке 5 представлена часть результатов работы алгоритма – дендрограммы, полученной на последнем этапе его работы.

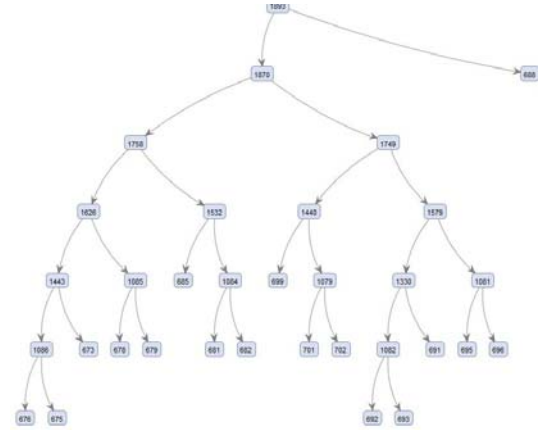


Рисунок 5 Работа алгоритма иерархической кластеризации

В результате применения предложенного подхода было получено окончательное решение, состоящее из двух кластеров, характеризующих поведение клиентов при осуществлении платежей, обладающее достаточно высокой точностью, согласующееся с экспертным мнением (см. Рис. 6). Из рисунка видно, что точность предложенного ансамбля превышает точность стандартного алгоритма К-средних, с различными метриками.

Рисунок 6 Сравнение точности алгоритмов

5 Заключение

Задача интеллектуального анализа и обработки Больших Данных последние несколько лет является предметом изучения множества специалистов и важной составляющей этого анализа указывается кластеризация этих данных, позволяющая приблизиться к решению проблемы трех V (объема данных для хранения - Volume, скорости обработки - Velocity и разнообразия исходных типов данных - Variety) [5]. Таким образом, кластерный анализ становится одним из ключевых в сферах обработки интенсивных данных, так как это один из

эффективных методов, который существует на сегодняшний день. Применяя ансамбль алгоритмов кластеризации, можно повысить достоверность разбиения данных на группы. Существенным является то, что данный метод может применяться в различных областях. Рассмотренный в данной статье ансамбль алгоритмов кластеризации нивелирует недостатки метрик расстояний для алгоритмов K-средних, тем самым повышая достоверность разбиения. Дальнейшее исследование предложенного ансамбля алгоритмов на основе меняющейся метрики расстояний планируется в рамках гранта РФФИ № 15-07-08742.

Литература

[1] J. Ghosh, A. Acharya Cluster ensembles. Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery. 2011. V. 1(4). P.305−315.

[2] Kausar Noreen;Abdullah Azween; Samir Brahim Belhaouari;Palaniappan Sellapan;AlGhamdi Bandar Saeed;Dey Nilanjan. Ensemble Clustering Algorithm with Supervised Classification of Clinical Data for Early Diagnosis of Coronary Artery Disease.// Journal of Medical Imaging and Health Informatics, V. 6, Number 1, February 2016, P. 78-87.

[3] Predictive Analytics Platform | RapidMiner https://rapidminer.com/

[4] UCI Machine Learning Repository https://archive.ics.uci.edu/ml/datasets/default+of+credit+card+clients (26.06.2016)

[5] А. К. Горшенин, С.Я. Шоргин. Разработка информационной технологии интеллектуального анализа больших данных // Современные проблемы прикладной математики, информатики, автоматизации, управления: Материалы 3-го Международного научно-технического семинара (Севастополь, 9–13 сентября 2013). –М.:ИПИ РАН, 2013. С. 104–114.

[6] А.П. Кулаичев. Методы и средства комплексного анализа данных. – М.: Форум — Инфра-М, 2006. — 512 с.

[7] Б.О. Лялька, Антонова-Рафи Ю.В. Оценка эффективности кластеризационных алгоритмов. // Научные труды SWORLD, 2015, №2 (39), С. 25-29.

[8] В.Б. Бериков. Классификация данных с применением коллектива алгоритмов кластерного анализа // Знания-Онтологии-Теории (ЗОНТ-2015), 2015, С. 29-38

[9] В. Б. Бериков. Коллектив алгоритмов с весами в кластерном анализе разнородных данных // Вестн. Том. гос. ун-та. Управление, вычислительная техника и информатика, 2013, №2 (23), с 22-31.

[10] В.С. Киреев. Оценка результатов кластеризации при использовании различных критериев качества// Программные продукты и системы, 2009, №3, С. 36-39.

[11] Д.Е. Мозохин, В.А. Калягин. Сравнительный анализ алгоритмов кластеризации в сетях фондовых рынков // Алгоритмы, методы и системы обработки данных. 2015, 4(33), С. 73-90

[12] И. Демин. Концепция кластера в технологиях интеллектуального анализа данных// Риск: Ресурсы, Информация, Снабжение, Конкуренция, 2012, 1, С. 260-263.

[13] И. А. Пестунов, В. Б. Бериков, Ю. Н. Синявский. Сегментация многоспектральных изображений на основе ансамбля непараметрических алгоритмов кластеризации // Вестник СибГАУ, 2010, №5(31), С. 56-64.

[14] С. А. Суслов. Кластерный анализ: сущность, преимущества и недостатки // Вестник НГИЭИ. 2010. №1, С. 51-57.

[15] С.А. Батуркин, Е.Ю. Батуркина, В.А. Зименко, И.В. Сигинов. Статистические алгоритмы кластеризации данных в адаптивных обучающих системах // Вестник РГРТУ, 2010, № 1 (31), С. 82-85.

[16] С. Л. Подвальный, А. В. Плотников, А. М. Белянин. Сравнение алгоритмов кластерного анализа на случайном наборе данных // Вестник ВГТУ, 2012, Т.8 №5, С. 4-6.

Development of Ensemble of Clustering Algorithms Based on Varying Distances Metrics

Pyotr V. Bochkaryov, Vasiliy S. Kireev

Currently there is an active accumulation of big data in various information environments, such as social, corporate, scientific and other domains. Intensive use of big data in various fields stimulates the increased interest of researchers to the development of methods and means of processing and analyzing massive data volumes with significant variety. One of the promising areas in data intensive analytics is cluster analysis, which allows to solve such problems as: reducing the dimension of the original dataset, identifying patterns, etc. In this article, the authors propose an ensemble of clustering algorithms, consisting of the basic algorithm K-means, characterized by one parameter - the distance metric between objects. For the evaluation of performance of the designed ensemble the open data archive of UCI was used.

Non-stationary signal analysis with multicollinearity predictors

© Dmitriy Malakhov © Olga Krasotkina
Moscow State University
malahovdi7@gmail.com o.v.krasotkina@gmail.com

Abstract

In this paper we focused on non-stationary signal analysis in task of hedge fund investment portfolio management. Information about funds and assets is updated every day each second, so this area has many raw data for analysis. We review existing methods and propose special signal transformation method that fixes one of main disadvantage of current signal analysis methods, applied to estimation of investment portfolio strategy (multicollinearity predictors). In addition, we described how we planed to apply full steps of data analysis (such as data extraction, transformation, analysis, visualization) to considered task.

1 Introduction

There is a wide class of signals analysis problems on the axis of discrete argument $Y = (y_t, t = 1, ..., N)$ (usually time). In such tasks, smoothly varying parameter $X = (x_t, t = 1, ..., N)$ is required to estimate in all observation points, which takes values from some set $x_t \in X$ and forms a hidden process, it is usually considered as random process. The main example of the problem is an estimating the investment portfolio management strategy, which plays a huge role in the modern investment analysis [7].

The main idea of investment portfolio management strategy estimation is to determine the percentage of portfolio at each time point, based on known values of portfolio return and assets cost in the stock market [1]. This problem is very interesting for shareholders and investment companies, which want to know how the fund manages own assets. On the contrary, the fund tries to hide its strategy, but the information about own profitability index must be published every day (hedge fund information is available on finance.yahoo.com).

The hedge fund industry has grown rapidly over the past decade to almost $1 trillion in assets and over 8,000 funds [5]. In many cases, the available information on a hedge fund is a time series of daily/monthly/yearly returns and fund's assets cost during holding period. This data is used for time series analysis, so this area can be called data intensive, because data permanently changes over time, and it can be used for real-time data analysis. Returns are then analysed using a variety of multi-factor models in order to detect the volatility of the hedge fund strategy.

One of the most common and effective multi-factor models for analysis of investment portfolios, called Returns Based Style Analysis (RBSA), was developed by Sharpe [3]. In the RBSA model, constrained linear regression of a relatively small number of single factors, represented by the periodic returns of generic market indices, approximates the periodic return of a portfolio. Each of these factors represents a certain investment style or sector. In order to account for distribution changes in portfolios, Sharpe used a moving window of some predefined length, assuming that the structure of the portfolio is constant inside the window.

Dynamic model in which portfolio weights change with time was proposed as a generalization of the stationary RBSA model [5]. This approach, called Dynamic Style Analysis (DSA), consists of estimating a time-varying regression model of the observed time series of a portfolio's periodic returns and those of assets market cost. This problem is closely related with the necessity of choosing the appropriate level of volatility of results ("smoothness"), ranging from the full stationarity of instant regression models to their absolute independence of each other. In selecting the volatility level, authors used the "leave-one-out" principle [2].

In this paper, we focus on assets volatility problem. Assume fund has high turnover ratio and also there is a strong correlation between assets, so-called multicollinearity assets. If we try to solve this problem using DSA, and find optimal "smoothness" by standard "leave-one-out" procedure, we will fail, because we will get large values of "smoothness" parameters, which means that assets share distribution is constant at each time point. But it is incorrect, because fund has high turnover ratio and it is portfolio tend to change. There exist methods to solve multicollinearity problem in stationary regression: a priori information, dropping a variable(s), transformation and selection of variables, regularization and additional or new data [4]. However, it is not known how this methods operate on non-stationary regression. In this paper, we want to provide a method for dealing with multicollinearity in the case of non-stationary (time-varying) regression problem of recovering the investment portfolio management strategy.

Proceedings of the XVIII International Conference «Data Analytics and Management in Data Intensive Domains» (DAMDID/RCDL'2016), Ershovo, Russia, October 11 - 14, 2016

2 Current methods

Suppose we have T time points. At each point, there are known values y_t portfolio returns. It is expected that the portfolio invested in n assets. Also there is an assumption, that no resources have been received from the outside and withdrawn from the portfolio during holding period. Based on this assumption the portfolio return is defined as a linear combination of its component assets return [2]. Share distribution of portfolio assets at each time point is used as the coefficients of the linear combination (for hedge funds linear combination coefficients may be negative for certain assets, which indicate that the fund took a specified portion of the assets in the loan [8]). Finally, we have a non-stationary regression model:

$$y_t = \alpha_t + \sum_{i=1}^{n} \beta_t^{(i)} x_t^{(i)} + \varepsilon_t = \alpha_t + \boldsymbol{\beta}_t^T \boldsymbol{x}_t + \varepsilon_t \quad (1)$$

where $\boldsymbol{x}_t = (x_t^{(1)}, ..., x_t^{(n)})^T$ - fund assets return at time point t, $\boldsymbol{\beta}_t = (\beta_t^{(1)}, ..., \beta_t^{(n)})^T$ - assets share distribution at time point t, ε_t - additive noise with zero mean value and unknown variance.

In Returns Based Style Analysis (RBSA) model by William Sharpe [3], it is supposed that the share distribution of portfolio is constant at each time point, i.e. $\boldsymbol{\beta}_t = \boldsymbol{\beta} = (\beta_t^{(1)}, ..., \beta_t^{(n)})^T$. According to RBSA we have a standard stationary regression model. In general, this assumption is incorrect, and investment portfolio management strategy should be considered as non-stationary model, because there are funds with large turnover ratio, which means that fund totally change their assets during holding period.

Next method, also known as Dynamic Style Analysis (DSA) [5,9,10], considers assets share distribution changing at each time point. We want to find regression coefficients $\boldsymbol{\beta}_t = (\beta_t^{(1)}, ..., \beta_t^{(n)}, \alpha_t)^T$ in model (1), that squared residuals φt of the approximation of the goal variable would be as small as possible at each time moment:

$$\varphi_t(\boldsymbol{\beta}_t) = (y_t - \boldsymbol{\beta}_t^T x_t)^2 \to \min_{\boldsymbol{\beta}_t} \quad (2)$$

The key point of the DSA is that fractional asset weights are considered as a hidden process assumed a priori to possess the Markov property:

$$\boldsymbol{\beta}_t = \boldsymbol{V}_t \boldsymbol{\beta}_{t-1} + \varepsilon_t \quad (3)$$

where matrices V_t determine the assumed hidden dynamics of the portfolio structure, and ε_t is the vector white noise, non-stationary in the general case, whose squared norm is to be minimized. So there exists the "smoothness" condition, expressed by the criterion:

$$\gamma_t(\boldsymbol{\beta}_{t-1}, \boldsymbol{\beta}_t) = (\boldsymbol{\beta}_t - \boldsymbol{V}_t \boldsymbol{\beta}_{t-1})^T \boldsymbol{U}_t (\boldsymbol{\beta}_t - \boldsymbol{V}_t \boldsymbol{\beta}_{t-1})$$
$$\gamma_t(\boldsymbol{\beta}_{t-1}, \boldsymbol{\beta}_t) \to \min_{\boldsymbol{\beta}_{t-1}, \boldsymbol{\beta}_t}, t = 1, ..., T \quad (4)$$

where $\boldsymbol{U}_t = (\lambda_{ii})$ defines appropriate norm (for instance, Euclidean norm), λ_{ii} is a "smoothness" coefficients.

Equation (3) determines the state-space model of a dynamic system, while (1) plays the role of its observation model. Required estimations $\hat{\boldsymbol{\beta}}_t$ can be determined from criterion, combined (2) and (4) for all time points:

$$J(\boldsymbol{\beta}_1, ..., \boldsymbol{\beta}_T) = \sum_{t=1}^{T} (y_t - \boldsymbol{\beta}_t^T x_t)^2 +$$
$$+ \sum_{t=2}^{T} (\boldsymbol{\beta}_t - \boldsymbol{V}_t \boldsymbol{\beta}_{t-1})^T \boldsymbol{U}_t (\boldsymbol{\beta}_t - \boldsymbol{V}_t \boldsymbol{\beta}_{t-1})$$
$$\hat{\boldsymbol{\beta}}_1, ..., \hat{\boldsymbol{\beta}}_T = \underset{\boldsymbol{\beta}_1, ..., \boldsymbol{\beta}_T}{\operatorname{argmin}} J(\boldsymbol{\beta}_1, ..., \boldsymbol{\beta}_T) \quad (5)$$

The positive parameter $\lambda = \lambda_{ii}$ in matrix U_t is responsible for the level of smoothness of regression coefficients. Thus, the smoothness parameter λ balances the two conflicting requirements: to provide a close approximation of portfolio returns and, at the same time, to control the smoothness of asset weights β_t over time.

A commonly used measure of regression model fit is its coefficient of determination R^2, so we use it to select the best model. The parameter λ is being found by the "leave-one-out" principle [2]:

1. Delete on y_t from $y_1, ..., y_T$
2. Analyze remaining $N-1$ elements $y_1, ..., y_{t-1}, y_{t+1}, ..., y_T$, and find estimation of shared distribution at deleted point
3. Count R^2 for deleted return value
4. Retry step 1-3 for all $y_1, ..., y_T$
5. Select λ, that maximize summary R^2

3 Problem formulation

The considered DSA method works incorrectly on some hedge funds, for example, on funds with high turnover ratio. Turnover ratio is a measure of the fund's trading activity, which is computed by taking the lower of purchases or sales (excluding all securities with maturities of less than one year) and dividing by average monthly net assets. A low turnover figure (20% to 30%) would indicate a buy-and-hold strategy. High turnover (more than 100%) would indicate an investment strategy involving considerable buying and selling of securities [6].

If we apply DSA (with "leave-one-out" procedure) to hedge fund with high turnover ratio, we will get constant assets shared distribution at each time point. However, it is incorrect, because high turnover ratio means that distribution changes over time. So volatility problem does not solve in that case. Financial analytical company

Markov Processes International (MPI)[1] faced with this problem in work practice.

We proposed that DSA method fails because assets in portfolio are strongly correlated (multicollinearity problem). Therefore, the main aim of this article is to develop approach, which copes with correlated regressors in signal analysis problem on example of the investment portfolio task.

4 Signal transformation

From existing methods of solving multicollinearity problem, we chose signal transformation to another space (with lower dimension). This transformation has an economic explanation for the investment portfolio estimation problem. The main contribution to the portfolio management strategy makes only the most important part of all assets. Thus, if hedge fund has n assets, we transform it to m assets in another space based on special transformation method. Then DSA method is applied to modified assets with lower dimension. After estimation of shared distribution for transformed assets, we will return to original signal space, and will get the final estimation of assets shared distribution.

We chose special methods of feature transformation, similar with Principal Component Analysis. Each asset in portfolio is decomposed into the sum:

$$x_i = \sum_{j=1}^{m} c_{i,j}\tilde{x_j}, j = 1..T, i = 1..T \quad (6)$$

where $\tilde{x_j}$ - eigenvectors of covariance matrix $A = (a_{i,j} = x_i^T x_j, i, j = 1, ..., n)$, corresponding to m maximum eigenvalues, and $c_{j,i} = \tilde{x_j} x_i$.

Then original model (1) is transformed to the new model:

$$y_t = \beta_t^T x_t = \sum_{i=1}^{n} \beta_{i,t} x_{i,t} = \sum_{i=1}^{n} \beta_{i,t} \sum_{j=1}^{m} c_{i,j}\tilde{x}_{j,t} =$$
$$= \sum_{j=1}^{m} (\sum_{i=1}^{n} \beta_{i,t} c_{i,j})\tilde{x}_{j,t} = \sum_{j=1}^{m} \tilde{\beta}_{t,j}\tilde{x}_{j,t} = \tilde{\beta_t}^T \tilde{x}_t \quad (7)$$

Then we apply DSA to new model $y_t = \tilde{\beta_t}^T \tilde{x}_t$ and find out a transformed shared distribution $\tilde{\beta}_t$. After that, it is only needed to recover original shared distribution based on transformed distribution in lower dimension space. From (7) we know that $\tilde{\beta}_{t,j} = \sum_{i=1}^{n} \beta_{t,i} c_{i,j}$, where $t = 1..T, j = 1..m$. To find desired shared distribution that try to approximate known turnover ratio, we need to solve problem, using above equation as constraints:

$$(\sum_{t=2}^{T} \sum_{i=1}^{n} |\beta_{t,i} - \beta_{t-1,i}| - TO)^2 \to \min_{\beta_1,...,\beta_t} \quad (8)$$

where TO - hedge fund turnover ratio.

This task can be solved with subgradient method with the objective function:

$$f(\boldsymbol{\beta}) = f(\beta_{1,1}, ..., \beta_{1,n}, ..., \beta_{T,1}, ..., \beta_{T,n}) =$$
$$= (\sum_{t=2}^{T} \sum_{i=1}^{n} |\beta_{t,i} - \beta_{t-1,i}| - TO)^2 +$$
$$+ \lambda_1 \sum_{t=1}^{T} (\sum_{i=1}^{n} \beta_{t,i} - 1) +$$
$$+ \lambda_2 \sum_{t=1}^{T} \sum_{j=1}^{m} (\tilde{\beta}_{t,j} - \sum_{i=1}^{n} \beta_{t,i} c_{i,j}) \to \min_{\boldsymbol{\beta}} \quad (9)$$

where λ_1, λ_2 - regularization parameters.

The first regularization summand corresponds to constraint, that shared assets distribution must have unit sum. The second corresponds to contraints from formula (7).

Subgradient method incrementally solve (9) until method converges:

$$\boldsymbol{\beta}_{k+1} = \boldsymbol{\beta}_k - \alpha_k \bigtriangledown f(\boldsymbol{\beta}_k) \quad (10)$$

where k - current iteration, $\bigtriangledown f$ - subgradient of function, α_k - constant at each iteration, corresponds to the rate of convergence.

After the convergence of the subgradient method, we find assets shared distribution β_t at each time point in the original assets space. This distribution is a final estimation.

5 Experiments

It this section we present our experiments for applying proposed signal transformation methods to improve current methods of signal analysis to the problem of investment portfolio management estimation.

We used data from Markov Processes International to verify our hypothesis. This dataset consists of about 150 hedge funds and 10 assets during the 60 days, 2 and 5 years. Assets correspond to the main sectors of S&P 500 index: Energy, Materials, Industrial, Consumer Discret, Consumer Staples, Health Care, Financials, Inf Technology, Telecomm Svcs, Utilities. In addition, a turnover ratio is known for each fund in sample, and for most of them, it is high.

The main characteristic of given data is a strong correlation between assets. Figure 1 illustrate assets in 60 days, in Figure 2 shows correlation assets heatmap.

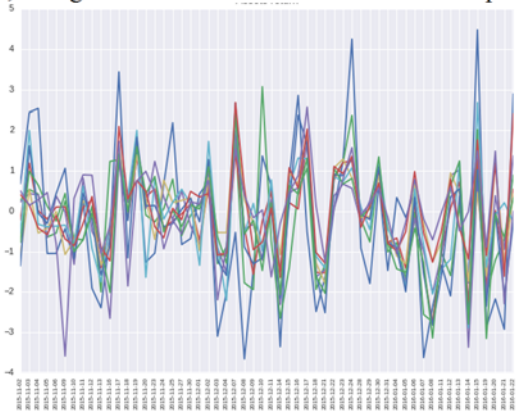


Figure 1 10 assets from MPI in 60 days.

[1] http://www.markovprocesses.com

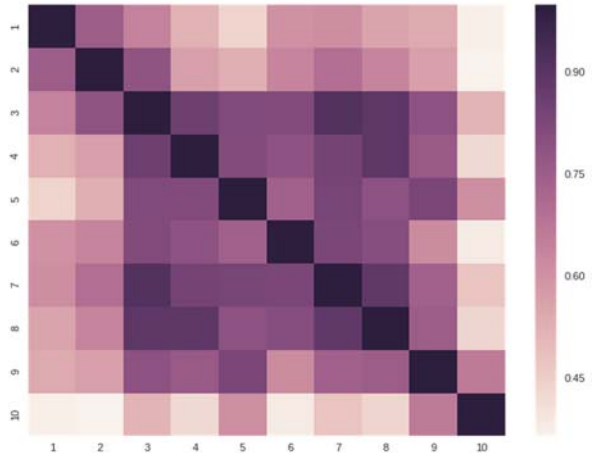


Figure 2 Correlation heatmap for 10 assets from MPI in 60 days.

As shown in Figure 1, Figure 2 considered S&P500 assets are strongly correlated with each other. If we apply standard DSA method to these funds based on given assets, we will get constant assets shared distribution at each time point, but it will be incorrect because considered hedge funds have high turnover ratios.

So assets first transform using method from previous section to the lower dimension space and then apply DSA method to transformed assets. Different values were applied for dimension, from 2 to all number of assets. We used for example one of hedge fund, named "AB Core Opportunities A". The predicted error of that hedge fund return shown in Figure 3. This transformation reduce predicted error with comparison assets in original space.

Figure 3 shows, that best way is to transform assets to space with dimension $m = 5$. We can see the comparison of given and predicted hedge fund return in Figure 4.

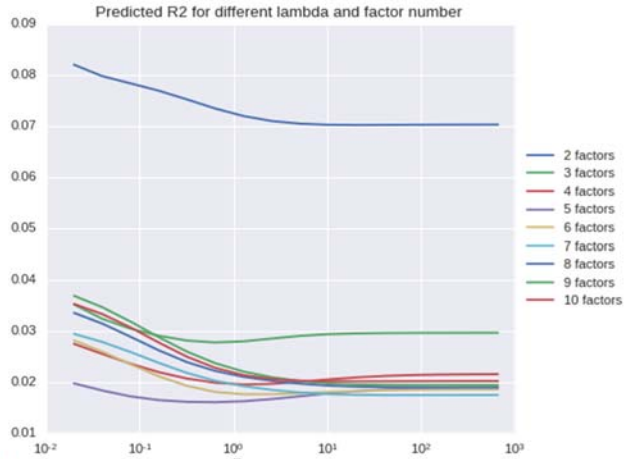


Figure 3 Predicted R^2 in transformed space for different values for dimension.

Our main goal is to estimate the assets share distribution at each time point and solve the hedge fund volatility problem. Except good predict hedge fund return at each time point we must estimate assets shared distribution. With purposed signal transformation method we find a good estimation of shared distribution similar with real distribution (based on root mean square error). It is shown in Figure 5, that estimated distribution in transformed space changes with time, so it corresponds to a high turnover ratio for hedge fund. Using this estimation original shared distribution, which corresponds to real investment portfolio management strategy, may be recovered.

Figure 4 Comparison original (blue color) and estimated hedge fund return (green color).

Therefore, using signal transformation before DSA method help to solve volatility problem of investment portfolio. This transformation cope with correlated signal regressors and help to apply original DSA method to transformed assets.

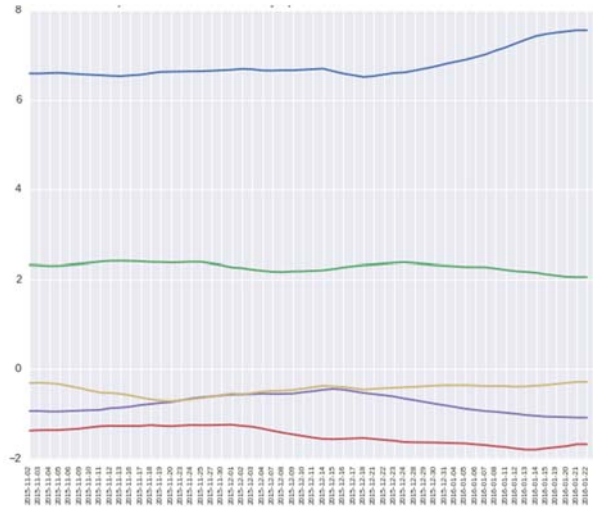


Figure 5 Top shared distribution components in transformed space.

5 Conclusions

In this paper, we focused on signal analysis with multicollineartiry regressors based on the investment portfolio strategy estimation. Original signal analysis methods work incorrect on hedge funds with high turnover ratio and predict constant assets shared distribution, but it must changes during time, because high turnover ratio.

We applied special signal transformation method similar with Principal Component Analysis and transformed original assets to lower dimension space. Then we applied standard DSA method to transformed assets and estimated shared distribution in that space. After we transformed it back and found final estimation of assets shared distribution at each time point. That

distribution is very similar with original distribution of assets at each time point, which is known from real data. Also predicted hedge fund return has low error, which means that estimated shared distribution successfully recover hedge fund investment portfolio management strategy.

6 Further work

In further work we would like to analyze deeper signals recovery methods based on investment portfolio management problem. We also want to apply all steps of data analysis to time-varying signals on the hedge funds as example.

Because this area data intensive, we need special technologies to cope with that task. It is supposed to use the following architecture to analyze the strategy of hedge funds, which is shown in Figure 6.

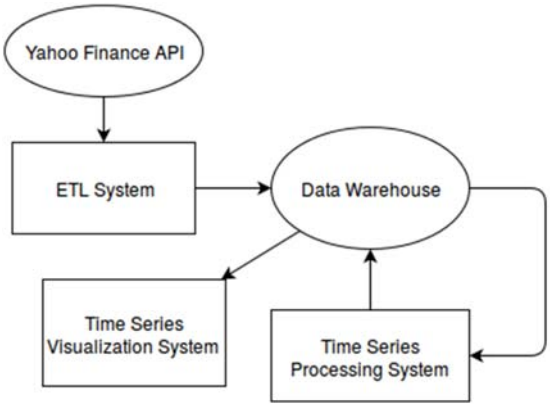


Figure 6 Architecture of hedge fund analyzing system.

We plan to use Yahoo Finance API as data source of time series. It includes frequently updated information about all mutual, hedge funds and all assets, also it includes historical data about each fund. Yahoo Finance provides JSON API, which is useful as a data source for hedge funds.

The next step is extract information from Yahoo Finance, transform it and load to the Data Warehouse. ETL System will receive tasks for download data and then store it to the data warehouse. Data warehouse and ETL System will be interact using Apache Kafka2 or Apache Storm3 which allow real-time streaming data.

We suggest to use distributed warehouse, for example, Hive4 or HBase5. We need distributed warehouse, because there are big amount of assets and funds in Yahoo Finance. Funds historical data will be stored for the last five years, because often time series analysis methods look at the very long time period.

In addition, we plan to develop special time series processing system on proposed data warehouse. It will be based on Spark6, which can quickly analyze and process big amount of data, and good integrated with Hive and HBase. This system will recover investment portfolio management strategy based on methods considered in this paper, also it will be used proposed signal transformation method to strongly correlated assets.

The final step of proposed architecture is visualization of recovered investment portfolio strategy. For it purpose we will use Tableau7, which supports the ability to visualize large and complex data.

References

[1] Krasotkina O.V. Algorithms of non-stationary signals models estimation with constraints. Candidate physical and mathematical sciences dissertation, 05.13.17. Tula, 2003

[2] Markov M.R., Mottl V.V., Muchnik I.B. Investment portfolio dynamic analysis: a new class of problems and methods of signal processing.

[3] Sharpe W. F.Determining a fund's effective asset mix. Investment Management Review, Vol. 2, 1988, No. 6 (December), pp. 59-69.

[4] Gujarati D. Multicollinearity: what happens if the regressors are correlated? Basic Econometrics (4th ed.). McGraw-Hill. pp. 364-369.

[5] Markov, M., Muchnik, I., Krasotkina, O., & Mottl, V. (2006, October). Dynamic analysis of hedge funds. In *The 3rd IASTED International Conference on Financial Engineering and Applications, ACTA Press, Cambridge*.

[6] Morningstar Investing Glossary http://www.morningstar.com/InvGlossary/turnover_ratio.aspx

[7] Reilly, F. K., & Brown, K. C. (2011). *Investment analysis and portfolio management*. Cengage Learning.

[8] Fung W., Hsieh D. A. Hedge fund benchmarks: A risk-based approach //Financial Analysts Journal. – 2004. – Т. 60. – №. 5. – С. 65-80.

[9] Markov M. et al. Machine-Learning for Dynamic Reverse Engineering of Hedge Funds //Machine Learning and Cybernetics, 2007 International Conference on. – IEEE, 2007. – Т. 5. – С. 2805-2812.

[10] Markov M., Mottl V., Muchnik I. Dynamic Style Analysis and Applications //Available at SSRN 1971363. – 2004.

[2] http://kafka.apache.org/
[3] http://storm.apache.org/
[4] http://hive.apache.org/
[5] http://hbase.apache.org
[6] http://spark.apache.org/
[7] http://www.tableau.com/

Разработка ансамбля алгоритмов классификации с использованием энтропийного показателя качества для решения задачи поведенческого скоринга

© И. А. Кузнецов © В. С. Киреев

Национальный исследовательский ядерный университет «МИФИ»,
Москва, Российская Федерация
iakuznetsov@mephi.ru vskireev@mephi.ru

Аннотация

С увеличением объёма цифровой информации в мире возрастает актуальность задачи фильтрации и обработки таких данных. С целью выявления действительно необходимой и полезной информации для пользователя, применяются подходы, основанные на принципах классификации объектов и отнесения исходного объекта к группе наиболее похожих на него. Основой для классификации выступают алгоритмы машинного обучения, а сама классификации успешно применяется в различных областях интенсивного использования данных, в частности, в рекомендательных системах. Представленная статья посвящена описанию разработанного ансамбля алгоритмов классификации при построении рекомендательных систем в области интеллектуального анализа данных. В работе представлены результаты исследования при формировании ансамбля алгоритмов для скоринговых систем с использованием слабоструктурированных данных, а предложенный ансамбль был протестирован на открытых данных портала UCI.

Введение

Одной из наиболее распространенных задач анализа данных является задача классификации. Задача классификации относится к разделу машинного обучения, который называется «обучение с учителем» (Supervised learning) [13]. Классификатором называется алгоритм, определяющий, какому из предопределенных классов принадлежит предъявляемый объект по вектору признаков. Подобный подход часто применяется в автоматизированных системах поддержки принятия решений таких, как рекомендательные системы, экспертные системы и т.д. [15].

Труды XVIII Международной конференции DAMDID/RCDL'2016 «Аналитика и управление данными в областях с интенсивным использованием данных», Ершово, 11-14 октября 2016

С ростом объемов данных, старые методы и способы обработки остаются в прошлом. Из-за обилия цифровой информации, поиск тематических статей или иных источников отнимает все больше времени и превращается в рутину. Зачастую, обработка данных должна выполняться в режиме онлайн, а требования к производительности и скорости работы являются довольно высокими.

Один из способов борьбы с такой проблемой являются рекомендательные и экспертные системы [12]. В своем ежегодном обзоре потенциала новых технологий и веяний, компания Gartner отмечает потенциал таких вещей, как умный советник, продвинутая аналитика с персональной доставкой информации, ответы на естественном языке, виртуальный персональный ассистент и прочее (см. рис.1[2]).

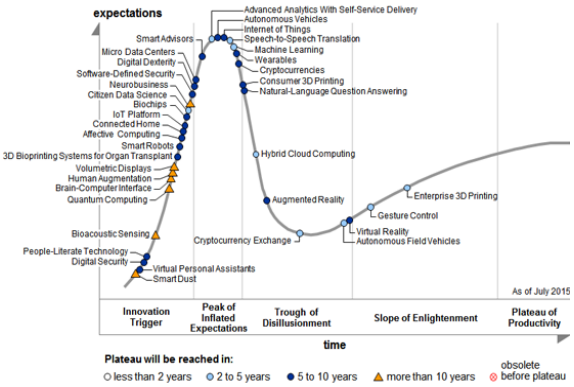


Рисунок 1 Цикл перспективных технологий

Использование описанных подходов нашло свое применение во многих областях, не исключая и финансовый сектор. Рост потребления благ человеком вынуждает в некоторых случаях использовать заемные средства. Любой заемщик для кредитной организации представляет определенный риск, а ее главной задачей является снижение такого риска. Появление систем оценки платежеспособности, основанное на численных статистических методах и дополненное средствами и инструментами машинного обучения, является одним из способов снижения риска для финансовой организации и роста ее доходов. Когда речь идет о больших суммах, рост точности прогнозного

значения даже на 1% может принести финансовой организации значительный доход.

Источниками данных для таких систем выступают статистические данные банков и иных кредитных организаций о выполнении клиентами своих обязательств. По каждому клиенту собирается и обрабатывается информация о его зарплате, имеющихся активах, образовании, кредитной истории, платежах по текущему кредиту и прочие данные.

Помимо описанных параметров, для оценки кредитоспособности потенциального клиента могут использоваться данные, не имеющие четкой структуры – слабоструктурированные данные. Такие данные могут быть представлены в различных форматах (как текстовом, так и графическом) и входят в общий перечень документов для оценки платежеспособности человека. К таким данным можно отнести различного рода рекомендательные письма, отзывы, справки и иные документы в свободной форме.

Помимо документов, которые предоставляет клиент, существуют и базы кредитных историй, которые доступны всем финансовым организациям для оценки заемщика. Ввиду различных стандартов кредитных организаций, некоторые поля могут заполняться в свободной форме: назначение кредита (как уже полученного, так и запросы), обеспечение, причина отказа и иные поля.

Целью данной статьи является представление авторского подхода к формированию ансамбля алгоритмов. В качестве прикладной области применения задачи классификации будут рассмотрено практическое использование ансамбля алгоритмов для формирования прогноза в области кредитного скоринга, где важную роль играют именно аналитические и экспертные системы принятия решений.

1 Алгоритмы классификации

Одним из подходов к решению задачи классификации является усиление простых классификаторов путём комбинирования примитивных слабых классификаторов в один сильный. Под силой классификаторов в данном случае подразумевается эффективность (качество) решения задачи классификации [11]. Основная идея использования ансамблей классификатор идентична той, когда при принятии важного решения человек пытается получить несколько различных мнений о своей проблеме, и на этой основе принимать решения.

Для оптимизации решения задач классификации и повышения точности работы алгоритмов выделяют несколько областей для исследования (см. рис.2).

Первым подходом для улучшения результата классификации является использование различных алгоритмов, отличных по своей природе и происхождению [14]. Существует значительное большое количество алгоритмов классификации, использование которых могут давать совершенно различные результаты.

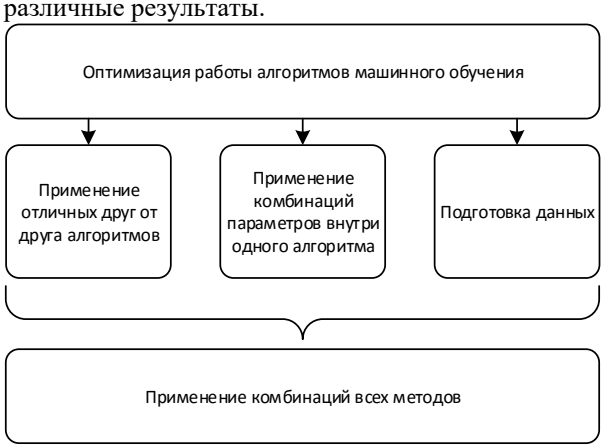


Рисунок 2 Методы оптимизации алгоритмов

Вторым подходом является оптимизация различных параметров известных алгоритмов, которые могут существенно влиять на результат. Например, в алгоритме RandomForest, такими параметрами являются [16]:

- количество деревьев входящих в алгоритм;
- минимальное число листов в дереве;
- минимальное расщепление в узле;
- минимальная и максимальная глубина;
- и другие.

В программной реализации алгоритма RandomForest число таких параметров значительно увеличивается за счет технических особенностей реализации и может доходить вплоть до 15.

Третьей подходом является работа с самими данными, которые подаются на вход алгоритма классификации. Подготовка данных – отдельная область, которая подразумевает обработку, очистку и приведение данных к машиночитаемому виду. Полученные данные могут быть изменены по какому-то признаку, могут быть добавлены новые значения, ранее не содержавшиеся в исходном наборе данных [8]. В некоторых случаях, данные могут быть удалены, т.к. не несут в себе полезной информации [4]. На основе исходных данных могут быть получены производные значения. Особенно это касается тех случаев, когда данные анонимны, т.е. представлены в виде цифровых значений или целиком в зашифрованном виде.

Таким образом, задача разработки оригинального алгоритма, который может являться ансамблем из нескольких известных алгоритмов, с различными весами и параметрами, и работающих со слабоструктурированными данными, является особо актуальной. При использовании данного подхода стоит обращать внимание на то, чтобы в итоговом ансамбле классификаторов были алгоритмы, имеющие разную природу происхождения [1]. Иначе, набор одинаковых алгоритмов будет выдавать схожие ответы и общее качество классификации значительно улучшить не получится.

Выделяют несколько способов для формирования ансамбля алгоритмов [6]:

- голосование большинством;
- веса пропорционально точности;
- использование условной вероятности;
- сероятностная формула Байеса;
- уменьшение дисперсии;
- независимость параметров друг от друга;
- взвешивание энтропии;
- плотностное взвешивание;
- и другие.

На сегодняшний день существует довольно много работ посвященных созданию скоринговых систем и различных комбинаций алгоритмов, позволяющих снизить риск невозврата кредита на основе данных о клиенте. Помимо широко распространённых и известных алгоритмов классификации, таких как Bagging и Boosting, создаются различные ансамбли алгоритмов классификации с предварительной кластеризацией [3], нечеткой логистической регрессией [9], а также использование нейросетей [7]. Представленные работы подробно описывают процесс, предшествующий принятию решения о выдаче кредита. В этом случае, используемые наборы данных в своей структуре содержат только количественные, категориальные и бинарные переменные.

Однако среди документов могут быть представлены и другие типы – текстовые данные. К ним могут относиться различного рода справки, анкеты, рекомендации и прочие неструктурированные текстовые данные. Текстовые данные имеют иную природу происхождения и требуют особого подхода к обработке и классификации.

Скоринговые системы направлены на работу с клиентом непосредственно до подписания договора, а затем работа скоринговой системы фактически заканчивается. В таком виде описанные системы никак не затрагивают процесс сопровождения клиентов и отслеживания рисков невозврата уже после получения кредита. Однако выделяют тип кредитного скоринга, который направлен на решение данной проблемы. Такой тип можно назвать «поведенческим», а основной задачей – регулярное отслеживание клиента в течение всего времени действия кредитного договора [5]. В качестве развития концепции о «поведенческой» составляющей клиента можно использовать не только финансовую информацию, но и оперировать ситуационными данными о клиенте, учитывать эмоциональное состояние клиента и прочее. Источником данных могут послужить социальные сети и иные открытые интернет источники.

Таким образом, работа с неструктурированными текстовыми данными может способствовать снижению потенциального риска выдачи кредита, а также учитывать в операционной деятельности кредитной организации и вероятность невозврата денежных средств в течение всего времени действия договора.

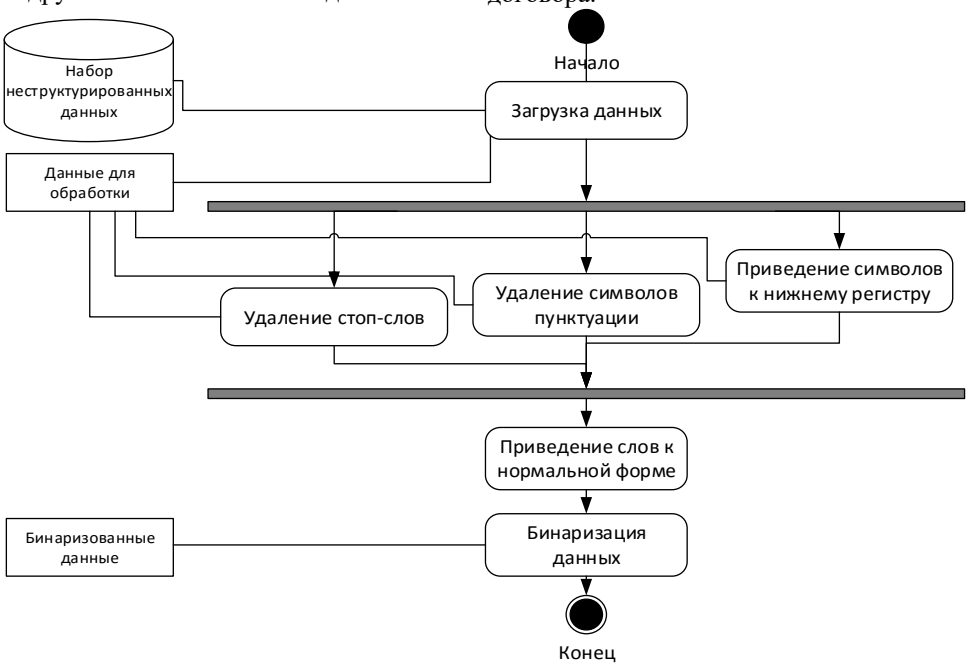


Рисунок 3 Предобработка текстовых данных

2 Ансамбль голосующих алгоритмов

В данной статье предлагается использовать следующие комбинации методов: использование условной вероятности и взвешивание энтропии.

В качестве основного коэффициента выступает энтропия, т.е. некая мера однородности результатов предсказания каждого алгоритма по отношению к правильному результату. Данная мера позволяет оценить качество работы конкретного классификатора для каждого класса.

Мера однородности рассчитывается по формуле (см. формулу 1):

$$H(x) = \sum_{i=1}^{n} p(i) \log_2 p(i) \qquad (1)$$

В качестве базовых алгоритмов могут рассматриваться любые простые алгоритмы, которые имеют различную природу происхождения. Это может быть комбинация алгоритмов «Случайный лес» (RandomForest - RF), «Наивный Байес» (Naïvebayes – NB), k-ближайших соседей (kNearestNeighbors – kNN) и других.

Также можно выделить классические ансамбли, которые зарекомендовали себя качеством и скоростью своей работы. К ним можно отнести классификаторы AdaBoost, Bagging и Boosting. Каждый из указанных ансамблей также можно встроить и использовать в других ансамблях.

Учитывая тот факт, что планируемый набор данных для классификации подразумевает работу со слабоструктурированными данными, то дополнительным шагом перед применением ансамбля алгоритмов является предобработка и выделение смысловой составляющей из слабоструктурированных данных.

На шаге предобработки загружаются неструктурированные данные, имеющие текстовый вид. Загруженные данные очищаются, нормализуются и приводятся к матричному виду. Алгоритм с предобработкой изображен на рисунке выше (см. рис.3).

Этап классификации состоит из трех шагов: обучение, тестирование, выбор результата (см. рис. 4).

На начальном этапе проводится обучение выбранных алгоритмов на основе обучающего набора данных. Выходными параметрами будут являться обученные модели по каждому классификатору.

Затем, на тестовой выборке проводится проверка обученных классификаторов и оценивается правильность полученных результатов. По каждому классификатору считается количество правильных и неправильных ответов. Для каждого алгоритма строится соответствующая матрица ошибок (см. матрицу 1):

$$\begin{pmatrix} n_{11} & n_{11} & \dots & n_{M1} \\ n_{21} & n_{11} & \dots & n_{M1} \\ \dots & \dots & \dots & \dots \\ n_{M1} & n_{M1} & n_{M1} & n_{MM} \end{pmatrix} \qquad (1)$$

Где n_{ij} – количество объектов, относящихся к i-ому классу, но классифицированных как класс j.

На основе представленной матрицы рассчитывается показатель энтропии для каждого алгоритма j по каждому классу i (см. матрицу 2):

$$\begin{pmatrix} H_{11} & H_{12} & \dots & H_{1N} \\ H_{21} & H_{21} & \dots & H_{2N} \\ \dots & \dots & \dots & \dots \\ H_{M1} & H_{M1} & \dots & H_{MN} \end{pmatrix} \qquad (2),$$

где H_{ij} – энтропия, M – число классов, N – число алгоритмов.

На следующем шаге выполняется расчет итогового класса на основе полученных значений. На тестовых данных считается вероятность по каждому алгоритму для каждого класса, затем полученные значения поочередно умножаются на соответствующую энтропию этого класса (при этом энтропия вычитается из единицы) (см. формулу 2).

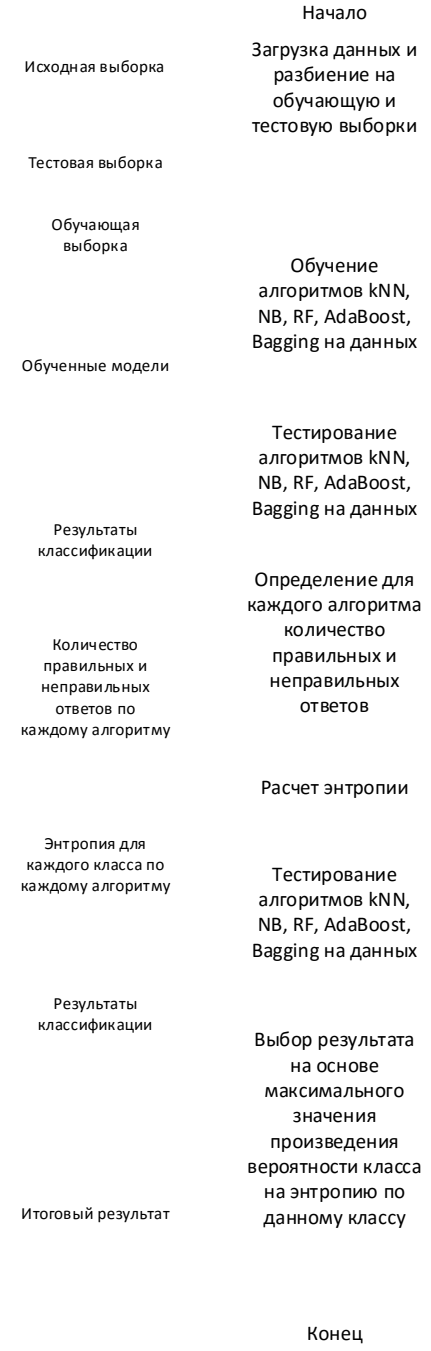


Рисунок 4 Алгоритм обучения

$$K = (1\text{-}p)*H \qquad (2),$$

где p – вероятность, H – энтропия, K – класс.

Затем выбирается класс с максимальным значением показателя качества K среди всех алгоритмов. Процедура проделывается для каждого

объекта выборки, до тех пор, пока не будет просмотрен весь список.

3 Тестирование алгоритма

В качестве предметной области для проверки результатов работы алгоритма был выбран набор данных, содержащих перечень кредитных платежей по клиенту в течение определенного периода времени (портал открытых данных UCI [10]). Представленный набор содержит базовую информацию о клиенте, получившем кредит: сумма, пол, образование, семейный статус и возраст. Помимо этого представленный набор данных содержит историю по платежам и суммам за полгода, что может позволить в дальнейшем сформировать некоторые паттерны для «поведенческого» типа скоринговых систем.

Задача: обучить классификатор и рассчитать точность его работы на тестовой выборке. Размер выборки составляет 30 тысяч различных записей. Количество классов в выборке равно двум: факт оплаты и факт неоплаты по кредитным обязательствам. Количество признаков равно 23. В качестве базовых алгоритмов были использованы: RandomForest, NaiveBayes, kNearestNeighbors, AdaBoost и Bagging.

После обучения базовых алгоритмов на 70% записей от исходного объема, и проведения тестирования на оставшихся 30% данных, была проведена серия экспериментов и были получены следующие результаты (см. таб. 1-4):

Таблица 1 Сравнение точности алгоритмов (первая итерация)

Алгоритм	Параметры	Точность
kNearestNeighbors	Кол-во соседей: 5	74,62%
RandomForest	Кол-во деревьев: 50	81,27%
NaiveBayes	Распределение: Бернулли	76,73%
AdaBoost	Кол-во классификаторов: 50	81,21%
Bagging	Кол-во классификаторов: 10	80,18%
Предлагаемый ансамбль		83,03%

Таблица 2 Сравнение точности алгоритмов (вторая итерация)

Алгоритм	Параметры	Точность
kNearestNeighbors	Кол-во соседей: 15	76,58%
RandomForest	Кол-во деревьев: 100	81,46%
NaiveBayes	Распределение: Бернулли	76,73%
AdaBoost	Кол-во классификаторов: 100	81,33%
Bagging	Кол-во классификаторов: 50	81,42%
Предлагаемый ансамбль		83,12%

Таблица 3 Сравнение точности алгоритмов (третья итерация)

Алгоритм	Параметры	Точность
kNearestNeighbors	Кол-во соседей: 40	77,47%
RandomForest	Кол-во деревьев: 200	81,48%
NaiveBayes	Распределение: Бернулли	76,73%
AdaBoost	Кол-во классификаторов: 200	81,30%
Bagging	Кол-во классификаторов: 75	81,22%
Предлагаемый ансамбль		83,25%

Таблица 4 Сравнение точности алгоритмов (четвертая итерация)

Алгоритм	Параметры	Точность
kNearestNeighbors	Кол-во соседей: 75	77,36%
RandomForest	Кол-во деревьев: 300	81,44%
NaiveBayes	Распределение: Бернулли	76,73%
AdaBoost	Кол-во классификаторов: 300	81,30%
Bagging	Кол-во классификаторов: 100	81,27%
Предлагаемый ансамбль		83,07%

Из таблицы видно, что предложенный алгоритм показывает сравнительно высокую точность, превосходящую точность метода случайного леса RF, а также алгоритмов AdaBoost и Bagging.Визуальное представление данных результатов представление можно увидеть на графике (см. рис.5).

В качестве направлений улучшения разработанного ансамбля планируются дальнейшие эксперименты с различными способами предобработки данных. Кроме того, планируется исследование применимости данного алгоритма на текстовых данных в области Web Mining и обработки сообщений из социальных сетей с целью извлечения

информации, способствующей оценки рисков потенциального заемщика.

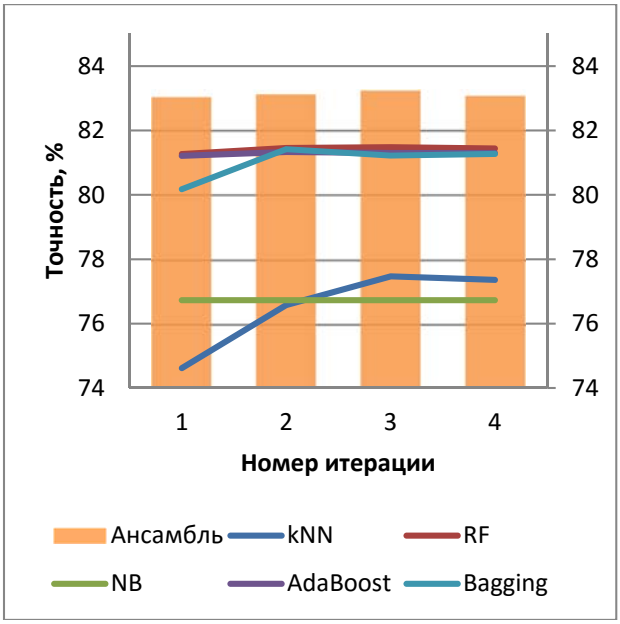


Рисунок 5 Сравнение точности предложенного ансамбля и классических подходов

4 Заключение

Решение задачи классификации востребовано во многих областях, связанных с обработкой больших объёмов данных, для поддержки процесса принятия решений. Существует множество алгоритмов классификации, эффективность работы которых ограничена объёмом и структурой данных, поэтому современный подход к данной проблеме заключается в конструировании ансамблей или комитетов более слабых алгоритмов. В данной работе предложен новый вариант ансамбля, использующий энтропийную меру в качестве меры однородности. Проведён эксперимент на открытых данных, который позволяет сделать заключение о перспективности предложенного метода классификации. Дальнейшие исследования по доработке разработанного ансамбля алгоритмов и его тестированию на слабоструктурированных данных поддержаны финансированием в рамках проекта № 57614X068 Федеральной Целевой Программы «Исследования и разработки по приоритетным направлениям развития научно-технологического комплекса России на 2014-2020 годы».

Литература

[1] TG Dietterich Machine-learning research - Four current directions. AI MAGAZINE Том: 18 Выпуск: 4, 1997, с.: 97-136.

[2] Gartner's 2015 Hype Cycle for Emerging Technologies Identifies the Computing Innovations That Organizations Should Monitor. http://www.gartner.com/newsroom/id/3114217 (дата обращения: 22.05.2016)

[3] Hongshan Xiao, Zhi Xiao, Yu Wang. Ensemble classification based on supervised clustering for credit scoring. Appl. Soft Comput. 43: 73-86 (2016)

[4] John P. Cunningham, Yu Byron M Dimensionality reduction for large-scale neural recordings. NATURE NEUROSCIENCE. Том: 17, Выпуск: 11, 2014, с.: 1500-1509.

[5] Kenneth Kennedy, Brian Mac Namee, Sarah Jane Delany, M. O'Sullivan, N. Watson. A window of opportunity: Assessing behavioural scoring. Expert Syst. Appl. 40(4): 1372-1380 (2013)

[6] Lior Rokach. Ensemble-based classifiers. Springer Science+Business Media B.V., 2009.

[7] Petr Hájek, Vladimír Olej. Intuitionistic Fuzzy Neural Network: The Case of Credit Scoring Using Text Information. EANN 2015: 337-346

[8] Priti Gupta, Omdutt Sharma. – Feature selection: an overview. International Journal of Information Engineering and Technology (IMPACT: IJIET) ISSN(E): Applied; ISSN(P): Applied Vol. 1, Issue 1, Jul 2015, 1-12

[9] So Young Sohn, Dong Ha Kim, Jin Hee Yoon. Technology credit scoring model with fuzzy logistic regression. Appl. Soft Comput. 43: 150-158 (2016)

[10] UCI Machine Learning Repository https://archive.ics.uci.edu/ml/datasets/default+of+credit+card+clients (дата обращения: 25.06.2016)

[11] Yeh, I. C., & Lien, C. H. The comparisons of data mining techniques for the predictive accuracy of probability of default of credit card clients. Expert Systems with Applications, 36(2), 2009, p. 2473-2480.

[12] А.И. Гусева, В.С. Киреев, П.В. Бочкарёв, И.А. Кузнецов. Исследование алгоритмов многомерной классификации научных данных//Фундаментальные исследования. – 2015. – № 11(5). – С: 868-874.

[13] В.И. Донской Алгоритмические модели обучения классификации: обоснование, сравнение, выбор. Издательство «ДИАЙПИ», Симферополь, 2014

[14] М. П. Кривенко, В. Г. Васильев Методы классификации данных большой размерности. – М.: ИПИ РАН, 2013. 204 с.

[15] С.А. Филиппов, В.Н. Захаров, С.А. Ступников, Д.Ю. Ковалев. Организация больших объемов данных в рекомендательных системах поддержки жизнеобеспечения, входящих в состав глобальных платформ электронной коммерции. Институт проблем информатики ФИЦ ИУ РАН. Selected Papers of the XVII International Conference on Data Analytics and Management in Data Intensive Domains (DAMDID/RCDL 2015), С. 119-124

[16] С.П. Чистяков Случайные леса: Обзор. Труды Карельского научного центра Российской академии наук, № 1, 2013, с. 117-136

Development of an ensemble of classification algorithms using the entropy quality measure for solving the problem of behavioral scoring

Igor A. Kuznetsov, Vasiliy S. Kireev

With increase of the amount of digital information the importance of a task of filtering and handling of such data increases in the world. For the purpose of identification of really required useful information for the user, the approaches based on the principles of classification of objects and rating of initial object to a group of the most similar to it are applied. Algorithms of machine learning act as a basis for classification, and classification itself is successfully applied in various data intensive areas, in particular, in the recommender systems. This article is devoted to the description of development of an ensemble of classification algorithms in case of creation of recommender systems in the field of data mining. In this article, the results of research for the case of forming ensemble of algorithms for scoring systems with use of semistructured data are presented, and the offered ensemble has been tested on open data of the UCI portal.

Быстрый алгоритм совмещения ультразвуковых дефектограмм рельсового пути

© А.С. Жуков © О.В. Красоткина

МГУ им. М.В. Ломоносова,
г. Москва

zhukov.msu@gmail.com o.v.krasotkina@yandex.ru

© А.А. Маленичев © В.В. Сулимова

ФГБОУ ВО «ТулГУ»,

г. Тула

malenichev@mail.ru vsulimova@yandex.ru

Аннотация

Работа посвящена разработке быстрого алгоритма совмещения ультразвуковых дефектограмм рельсового пути по нескольким каналам. Актуальной проблемой в процессе анализа дефектограмм является восстановление пропущенных данных по предыдущим проходам. Данная задача уже была решена ранее [6], однако разработанный алгоритм отличается высокой вычислительной сложностью. В настоящей работе разработан быстрый алгоритм восстановления пропущенных данных, в рамках которого создана параллельная реализация метода динамического программирования на GPU. Анализ результатов показал, что достигнута достаточно хорошая производительность для быстрой обработки данных специалистами-дефектоскопистами.

1 Введение

Причиной аварий на железных дорогах зачастую являются дефекты железнодорожного полотна. Они могут появляться из-за плохих погодных условий, неправильной эксплуатации, а также по другим причинам. С целью снижения вероятности образования изломов и схода поездов с рельсов, необходим своевременный ремонт проблемных участков.

Для поиска неисправностей применяются различные средства контроля. При обычном осмотре специалистом рельса возможно выявление лишь малого числа типов дефектов, так как большинство из них располагаются во внутренней части рельса. Поэтому, в настоящее время для детектирования неисправностей в железнодорожном полотне наиболее широко применяются ультразвуковые дефектоскопы [9]. В данной работе рассматриваются дефектограммы, полученные с помощью ультразвукового дефектоскопа АВИКОН-11 УДС2-114 (далее по тексту – дефектоскоп).

Данные, полученные с дефектоскопа хранятся в специальном формате. Ультразвуковой зондирующий импульс посылается с частотой 1000 Гц по 7 независимым каналам в разных направлениях, просматривая рельс в различных сечениях. Регистрация ультразвуковой дефектограммы производится посредством записи отраженных сигналов. При этом регистрируются амплитуды и времена задержки отраженных импульсов в диапазоне, обеспечивающем просмотр рельса на всю его глубину. Так как скорость дефектоскопа непостоянна, то результаты зондирования располагаются на шкале времени обычно неравномерно [6]. Как правило в каждый момент времени для каждого канала мы имеем один импульс, характеризующийся амплитудой и задержкой сигнала. Однако, иногда возникают случаи, когда ультразвуковой сигнал не успевает затухнуть при отражении от дефекта и возникает ситуация множественного переотражения, поэтому регистрируется несколько импульсов с разной задержкой и амплитудой в одной точке контроля. В настоящее время все данные, полученные с устройства, обрабатываются вручную дефектоскопистами с помощью программы «Авикон 11».

Главной задачей при анализе ультразвуковых дефектограмм является распознавание дефектов. При этом необходимо представлять участки дефектограммы в виде, удобном для применения алгоритмов машинного обучения. Так как фрагменты дефектограммы, подлежащие анализу, содержат различное число отсчетов и имеют разную

Труды XVIII Международной конференции DAMDID/RCDL'2016 «Аналитика и управление данными в областях с интенсивным использованием данных», Ершово, 11-14 октября 2016

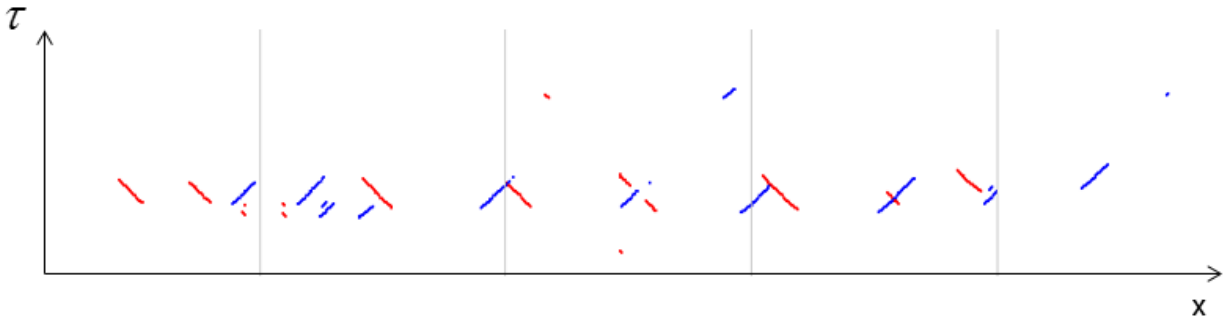


Рисунок 1 Пример дефектограммы, развертка типа B (2 канала – красный и синий).

длину, представить их непосредственно в пространстве действительнозначных признаков невозможно. В задачах подобного рода зачастую проще и удобнее ввести на объектах понятия расстояние и погрузить множество объектов в подходящее метрическое пространство.

Данная проблема уже была решена ранее в работах [6, 10]. Для подсчета расстояния в работах используется процедура совмещения различных проходов рельсового полотна. На первом этапе процедуры строится матрица попарных расстояний (матрица несходства) всех точек участков. На втором этапе с помощью метода динамического программирования находится оптимальное парное соответствие двух дискретных сигналов, полученных с дефектоскопа. Однако, проблема данного подхода состоит в достаточно маленькой скорости обработки данных на некоторых стадиях, а также встает проблема ограниченности оперативной памяти. Таким образом было решено разработать быстрый алгоритм совмещения ультразвуковых дефектограмм.

Основной проблемой является большая размерность данных: на 2 метра рельсового пути генерируется матрица несходства размером порядка 1 млн элементов (на 1 рельс), а обработка прямого хода метода динамического программирования занимает около 11 секунд. Учитывая протяженность железнодорожного полотна в РФ, нетрудно представить, как долго будет происходить обработка данных.

В настоящей статье данная проблема успешно решена: разработан параллельный алгоритм для процедуры динамического программирования. С целью ускорения вычислений была использована технология CUDA, которая доступна каждому владельцу современных видеокарт производства NVIDIA.

Статья имеет следующую структуру. В разделе 2 описывается процедура парного выравнивания, реализованная в рамках статьи [6]. В третьем разделе производится анализ скорости работы алгоритма (обработки данных), а также описывается схема распараллеливания алгоритма, описанного во второй части. В четвертом разделе производится анализ производительности предложенного решения. В разделе 5 делаются выводы, обобщающие полученные результаты.

2 Поиск оптимальных парных соответствий двух дефектограмм

2.1 Сравнение элементов двух дефектограмм

Существует множество методов сравнения дискретных сигналов, описанных в статьях [4, 3] и др. Однако, поскольку дефектоскоп движется по рельсам неравномерно (с различной скоростью), то дефектограммы, полученные с одного участка могут иметь разную длину. Поэтому, в предыдущих работах [6, 10] для установления оптимальных парных соответствий использовался метод динамической трансформации временной шкалы (Dynamic time warping) с нестандартной мерой несходства, который может работать с учетом этого фактора. Впервые этот подход был применен для распознавания речи, где использован для сопоставления двух разных сигналов с одной и той же произнесенной фразой. В классическом подходе в данном методе используется евклидова метрика. Метод состоит в следующем [2]: на первом этапе строится матрица попарных расстояний между элементами сигналов, на втором этапе с помощью метода динамического программирования строится матрица трансформаций (сжатие и растяжение сигнала), на третьем этапе строится оптимальный путь трансформации.

Дефектограмма есть двухкомпонентный дискретный сигнал, составленный из пар $(\tau_i, a_i) \in R^2$, $i = 1,...,n$, где τ_i – задержка, а a_i – амплитуда, n – число отражений зондирующего импульса в данный момент времени. Обычно, графически она представляется в виде развертки типа «B»: «время распространения ультразвуковых колебаний в рельсе — координата пути» (рис. 1) [8]. Каждый элемент дефектограммы представляет собой импульсный сигнал [10], в котором могут содержаться несколько импульсов, полученных из различных каналов. Таким образом, ввиду дискретности подобных сигналов сравнивать их напрямую не предоставляется возможным. Ранее в работе [6] был предложен метод представления сигналов в удобном виде для дальнейшего сравнения. Данная

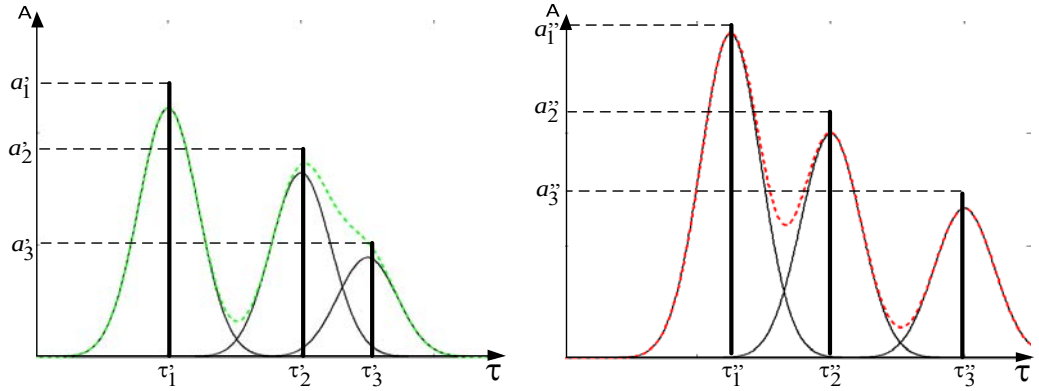


Рисунок 2 Графическое представление модели описания элемента дефектограммы по одному каналу в виде совокупности нормальных распределений

математическая модель основана на том, что каждый импульс сигнала по отдельному каналу описывается суммой нормальных распределений с математическим ожиданием, равным τ_i (задержка i-го импульса) и некоторой дисперсией σ.

$$f(\tau \mid x) = \sum_{i=1}^{n} a_i \frac{1}{\sigma\sqrt{2\pi}} \exp\left(-\frac{1}{2\sigma^2}(\tau - \tau_i)^2\right) \quad (2.1)$$

Использование такой математической модели обусловлено тем, что она позволяет получить меру несходства на участках дефектограммы, удовлетворяющую всем аксиомам метрики.

На рисунке 2 показана графическая интерпретация данной математической модели.

Пусть $\mathbf{x}' = (\tau_i', a_i') \in R^2$, $i = 1,...,n'$ и $\mathbf{x}'' = (\tau_i'', a_i'') \in R^2$, $i = 1,...,n''$ - два элемента дефектограммы. Меру их несходства будем вычислять как интеграл от модуля разности двух функций (2.1), что можно интерпретировать как площадь области несходства двух дефектограмм, ограниченной графиками распределений:

$$r(\mathbf{x}',\mathbf{x}'') = \sqrt{\int_{-\infty}^{\infty} \left[f(\tau \mid \mathbf{x}') - f(\tau \mid \mathbf{x}'') \right]^2 d\tau} \quad (2.2)$$

Интеграл (2.2) можно вычислить аналитически. В результате, меру несходства можно записать следующим образом:

$$r(\mathbf{x}',\mathbf{x}'') = \sqrt{\rho(\mathbf{x}',\mathbf{x}') + \rho(\mathbf{x}'',\mathbf{x}'') - 2\rho(\mathbf{x}',\mathbf{x}'')},$$

$$\rho(\mathbf{x}',\mathbf{x}'') = \sum_{i=1}^{n'} \sum_{j=1}^{n''} a_i' a_j'' \exp\left[-\left(\frac{\tau_i' - \tau_j''}{2\sigma}\right)^2\right]. \quad (2.3)$$

Если же мы сравниваем элемент $\mathbf{x}' = (\tau_i', a_i') \in R^2$, $i = 1,...,n'$ с некоторым нулевым элементом $z = (0,0)$, то вычислить данную меру можно по более простой формуле:

$$r(\mathbf{x}', z) = \sqrt{\sum_{i=1}^{n'} \sum_{k=1}^{n'} a_i' a_k' \exp\left[-\left(\frac{\tau_i' - \tau_k'}{2\sigma}\right)^2\right]} \quad (2.4)$$

На рисунке 3 представлена графическая интерпретация данной меры несходства.

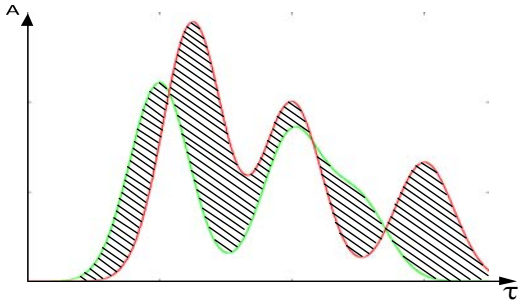


Рисунок 3 Графическое представление меры несходства двух элементов дефектограмм

Практика исследований показывает, что информации, зафиксированной ни по одному из отдельно взятых ультразвуковых каналов, оказывается не достаточно для решения задач анализа ультразвуковых дефектограмм с требуемой степенью точности [6, 10].

Совместное использование информации, полученной с разных каналов, в рамках данной работы осуществляется путем введения расширенной меры несходства элементов дефектограммы, представляющей собой линейную комбинацию частных мер несходства (2.3) с некоторыми неотрицательными весами $\alpha_i > 0$, в сумме составляющими единицу:

$$r(\mathbf{x}',\mathbf{x}'') = \sum_{i=1}^{m} \alpha_i \tilde{r}(x_i', x_i''), \quad \sum_{i=1}^{m} \alpha_i = 1, \quad (2.5)$$

где $\mathbf{x}'$ и $\mathbf{x}''$ – два m-мерных вектора, представляющих элементы дефектограммы.

Следует обратить внимание, что такой способ комбинирования не требует внесения изменений в общую схему сравнения дефектограмм и позволяет учитывать один или сразу несколько ультразвуковых каналов в зависимости от заданных коэффициентов линейной комбинации.

2.2 Поиск оптимальных парных соответствий

Под оптимальным парным выравниванием понимают установление парных соответствий между элементами сравниваемых сигналов в соответствии с

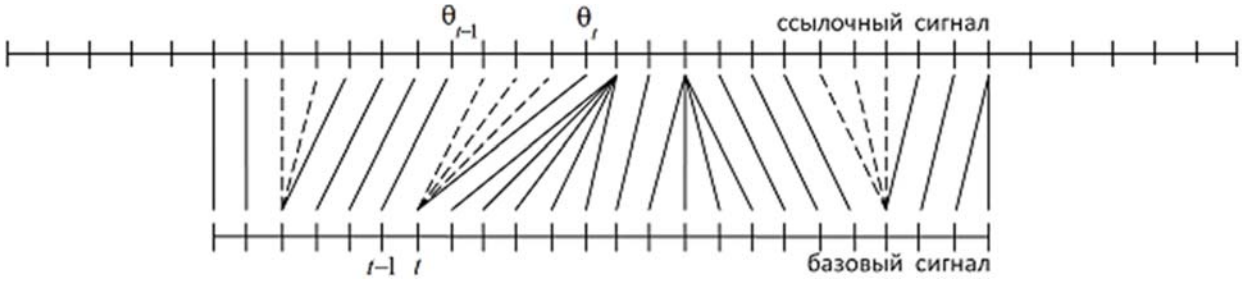


Рисунок 4 Пример расстановки ссылок при локально-глобальном выравнивании

некоторым критерием оптимальности. Процесс парного выравнивания обычно сопровождается локальной деформацией сравниваемых сигналов, т.е. локальным растяжением или сжатием их осей.

Сформулируем задачу парного выравнивания фрагментов дефектограмм для общего случая. Пусть мы имеем два требующих сравнения многокомпонентных дискретных сигнала, $\mathbf{X} = (\mathbf{x}_1,...,\mathbf{x}_{N_\mathbf{X}})$ и $\mathbf{Y} = (\mathbf{y}_1,...,\mathbf{y}_{N_\mathbf{Y}})$, представляющих фрагменты дефектограмм и имеющих в общем случае разную длину. Один из них – $\mathbf{X} = (\mathbf{x}_1,...,\mathbf{x}_{N_\mathbf{X}})$ – примем за «базовый», а другой – $\mathbf{Y} = (\mathbf{y}_1,...,\mathbf{y}_{N_\mathbf{Y}})$ – за «ссылочный».

Пусть также на множестве элементов этих сигналов определена метрика $r(\mathbf{x},\mathbf{y})$, например, согласно (2.5)

Требуется найти оптимальные парные соответствия элементов сравниваемых дефектограмм $\hat{T} = \arg\min_T J(\mathbf{X},\mathbf{Y},T)$, т.е. каждому элементу из базового сигнала сопоставить некоторый элемент из ссылочного, а результат запомнить в виде таблицы ссылок $T = (\theta_t, t = 1,..., N_\mathbf{X})$, где $\theta_t \in \{1,...N_\mathbf{Y}\}$ – абсолютная ссылка (номер элемента в ссылочном сигнале, соответствующий элементу t в базовом).

Очевидно, что решение данной задачи будет существенным образом зависеть от того, какой критерий парного выравнивания $J(\mathbf{X},\mathbf{Y},T)$ будет выбран для установления парных соответствий между элементами сравниваемых фрагментов дефектограмм. Будем применять критерий локально-глобального выравнивания [10]:

$$J(\mathbf{X},\mathbf{Y},T) = r(\mathbf{x}_1,\mathbf{y}_{\theta_1}) + \sum_{t=2}^{N_\mathbf{X}} \gamma(\theta_{t-1},\theta_t),$$

$$\gamma(\theta_{t-1},\theta_t) = \begin{cases} \beta + r(\mathbf{x}_t,\mathbf{y}_{\theta_t}), & \theta_t = \theta_{t-1}, \\ \beta\,|\,\theta_t - \theta_{t-1} - 1\,| + \sum_{j=\theta_{t-1}+1}^{\theta_t} r(\mathbf{x}_t,\mathbf{y}_j), & \theta_t > \theta_{t-1} \\ \infty, & \theta_t < \theta_{t-1} \end{cases} \quad (2.6)$$

где $\beta > 0$ – штраф на непараллельные ссылки, которые соответствуют локальным деформациям осей сигналов, представляющих фрагменты дефектограмм, при их выравнивании, а запись $r(\mathbf{x}_1,\mathbf{y}_{\theta_1})$ означает, что первый элемент базовой дефектограммы может соответствовать любому элементу ссылочной дефектограммы.

Данный критерий обеспечивает учет всех отсчетов как базового, так и ссылочного сигнала, добавляя дополнительные ссылки, связывающие t-й отсчет базового сигнала с отсчетами $\theta_{t-1}+1,...,\theta_t$ в случае, если $\theta_t > \theta_{t-1}$. Кроме того, данный критерий не допускает наличия так называемых «перекрестных ссылок», накладывая на них бесконечный штраф. Это означает, что, в соответствии с данным критерием, если $(t-1)$-й отсчет базового сигнала ссылается на θ_{t-1}-й отсчет ссылочного сигнала, то следующий t-й отсчет базового сигнала может ссылаться только на отсчет с большим номером $\theta_t \geq \theta_{t-1}$. Схематично идея парного выравнивания сигналов представлена на рисунке 4. Критерий (2.6) устроен таким образом, что он разрешает оставлять свободными начальные и конечные элементы именно ссылочного сигнала. Таким образом, при его использовании для поиска в длинном сигнале фрагмента, похожего на короткий сигнал, необходимо, чтобы именно базовый сигнал является более коротким.

2.3 Метод динамического программирования для локально-глобального парного выравнивания

Для сигналов $\mathbf{X} = (\mathbf{x}_1,...,\mathbf{x}_{N_\mathbf{X}})$ и $\mathbf{Y} = (\mathbf{y}_1,...,\mathbf{y}_{N_\mathbf{Y}})$ обозначим критерий (2.6) как $J(T) = J(\mathbf{X},\mathbf{Y},T)$ (для компактной записи). Функция (2.6) является сепарабельной [5], поскольку может быть представлена как сумма более простых функций, зависящих только от двух соседних переменных θ_{t-1} и θ_t. Экстремум этой функции ищется с помощью процедуры динамического программирования, осуществляемой в 2 прохода: прямой и обратный.

На прямом ходе осуществляется движение вдоль отсчетов базового сигнала и вычисление неполного критерия $\tilde{J}_t(\theta_t)$ по следующей рекуррентной схеме:

$$t = 1: \qquad \tilde{J}_1(\theta_1) = J(\theta_1), \quad \theta_1 \in \{1,..., N_\mathbf{Y}\}$$

$$t = 2,..., N_X: \ \tilde{J}_t(\theta_t) = \min_{\tau \in \{1,...,\theta_t\}} \left[\gamma(\tau,\theta_t) + \tilde{J}_{t-1}(\tau) \right], \quad \theta_t \in \{1,..., N_\mathbf{Y}\}.$$

Кроме того, запоминаются обратные рекуррентные соотношения:

$$t = 2,...,N_X: \quad \hat{\theta}_{t-1}(\theta_t) = \arg\min_{\tau\in\{1,...,\theta_t\}}\left[\gamma(\tau,\theta_t) + \tilde{J}_{t-1}(\tau)\right],$$
$$\theta_t \in \{1,...,N_{\mathbf{Y}}\}.$$

На обратном ходе осуществляется движение в обратную сторону вдоль отсчетов базового сигнала. При этом в каждом отсчете определяется оптимальное значение соответствующей ссылки:

$$\hat{\theta}_t = \arg\min_{\tau\in\{1...N_{\mathbf{Y}}\}}[\tilde{J}_{N_{\mathbf{X}}}(\tau)],\ \hat{\theta}_{t-1} = \hat{\theta}_{t-1}(\hat{\theta}_t),$$

для $t = N_{\mathbf{X}},...,2$.

Результатом выполнения процедуры является таблица ссылок $\hat{T}$, а также достигнутый минимум критерия

$$J(T) = \tilde{J}_{N_{\mathbf{X}}}(\hat{\theta}_{N_{\mathbf{X}}}),\ (2.7)$$

который может служить численной мерой несходства двух фрагментов дефектограмм.
Величину минимума $J(T)$ можно назвать расстоянием между двумя сигналами, если разделить ее на количество расставленных ссылок.

Представим алгоритм в виде ориентированного графа парных соответствий. В этом графе каждая вершина связывает некоторый элемент базового сигнала с определенным элементом ссылочного.

На рисунке 5 представлена графическая интерпретация: изображены возможные пути перемещения по графу парных соответствий для каждой вершины графа. В кружок обведена текущая вершина, жирным выделены точки, для которых значение критерия уже вычислено.

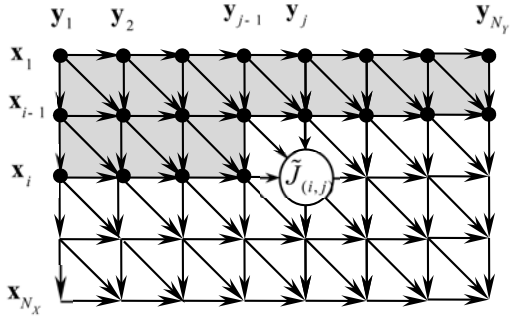


Рисунок 5 Иллюстрация алгоритма поиска оптимального выравнивания при помощи графа парных соответствий

Для такого представления любое локально-глобальное выравнивание может быть представлено в виде пути в этом графе, как показано на рисунке 6. При передвижении по горизонтальному ребру происходит локальное растяжение оси базового сигнала, по горизонтальному – ссылочного.

Суть работы алгоритма заключается в прохождении по вершинам графа и вычислению на каждой вершине неполного критерия $\tilde{J}_{(i,j)}$

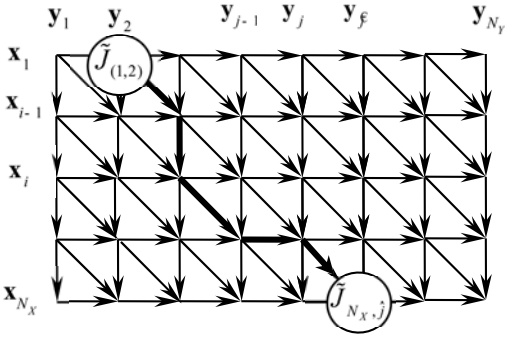


Рисунок 6 Пример локально-глобального выравнивания в виде оптимального пути на графе парных соответствий

$$\tilde{J}_{(1,j)} = r(\mathbf{x}_1, \mathbf{y}_j), \quad j = 1,...,N_Y,$$

$$\tilde{J}_{(i,j)} = r(\mathbf{x}_i, \mathbf{y}_j) + \min\begin{cases}\beta + \tilde{J}_{(i-1,j)},\\ \tilde{J}_{(i-1,j-1)},\\ \beta + \tilde{J}_{(i,j-1)},\end{cases} \quad (2.8)$$

$i = 2,...,N_X,\ j = 1,...,N_Y,$

Последняя вершина оптимального пути выбирается как вершина нижней грани, доставляющая минимум неполного критерия

$$\hat{j} = \arg\min_{j=1,..,N_Y} \tilde{J}_{(N_X,j)},$$

а его оптимальное значение соответствует значению в этой вершине:

$$J(T,X,Y) = \tilde{J}_{(N_X,\hat{j})}$$

Далее, на обратном ходе, происходит процедура восстановления траектории оптимального выравнивания (строится таблица соответствий элементов двух дефектограмм).

3 Анализ алгоритма парного выравнивания

3.1 Скорость работы алгоритма

Описанный алгоритм парного выравнивания является достаточно эффективным с качественной точки зрения. Однако, на практике было выяснено, что скорость его работы достаточно мала, особенно при больших объемах данных (а именно с ними и приходится сталкиваться на практике ввиду протяженности железных дорог в РФ).

Как было описано выше, алгоритм состоит из двух основных стадий:

1. Вычисление элементов матрицы несходства
2. Проход метода динамического программирования

Первый этап достаточно хорошо распараллеливается: нет никакой зависимости по данным, каждый элемент матрицы несходства можно вычислять в отдельном процессе. Однако, на преобразование данных, а также на пересылку будет тратиться больше времени, чем на сами вычисления.

Можно также заметить, что он занимает не так много времени, как, например, второй.

Второй этап, в отличие от первого работает значительно медленнее и, однозначно, требует параллельной реализации, для того чтобы эффективно использовать методику восстановления пропущенных данных на практике. Здесь уже имеется зависимость по данным, поэтому необходима разработка метода, позволяющего эффективно использовать вычислительные ресурсы.

3.2 Схема распараллеливания

На каждом шаге прямого хода метода динамического программирования считаются следующие соотношения:

$$\tilde{J}_{(1,j)} = r(\mathbf{x}_1, \mathbf{y}_j), \quad j = 1,..., N_Y,$$

$$\tilde{J}_{(i,j)} = r(\mathbf{x}_i, \mathbf{y}_j) + \min \begin{cases} \beta + \tilde{J}_{(i-1,j)}, \\ \tilde{J}_{(i-1,j-1)}, \\ \beta + \tilde{J}_{(i,j-1)}, \end{cases}$$

$$i = 2,..., N_X, \; j = 1,..., N_Y,$$

То есть, для подсчета очередного элемента $\tilde{J}_{(i,j)}$ должны быть известны (посчитаны) элементы $\tilde{J}_{(i-1,j)}, \tilde{J}_{(i-1,j-1)}, \tilde{J}_{(i,j-1)}$. Чтобы посчитать граничные элементы матрицы, добавляется дополнительная строка сверху и столбец слева элементов, равных бесконечности. Таким образом, чтобы посчитать, например, 1-й элемент 2-й строки матрицы, должен быть известен элемент, находящийся над ним, т.е. $\tilde{J}_{(1,1)}$. Далее, чтобы посчитать $\tilde{J}_{(2,2)}$, должен быть вычислен элемент $\tilde{J}_{(1,2)}$ ($\tilde{J}_{(1,1)}$ и $\tilde{J}_{(2,1)}$ уже будут вычислены в любом случае).

Таким образом, если обработку каждой строки матрицы запустить в отдельном процессе, то ее элементы могут быть вычислены в порядке, как показано на рисунке 7.

4 Реализация и анализ производительности

Существует множество различных способов реализации параллельных алгоритмов, таких как MPI, Open MP, Hadoop MapReduce, многопоточность на одном процессоре, использование GPU и др. В настоящей работе был реализован подход, использующий видеокарты NVIDIA™ для высокопроизводительных вычислений (CUDA). Данный подход был использован по ряду причин. Во-первых, технология является доступной для обычных пользователей ПК, в которых установлен графический процессор; она исключает проблему разворачивания дорогих высокопроизводительных систем. Во-вторых, при использовании кластерной архитектуры достаточно большое количество времени тратится на передачу данных. Поскольку настоящая задача предполагает помочь операторам быстро анализировать дефектограммы, и скорее будет использоваться для разовых задач, такое распараллеливание окажется неэффективным и бессмысленным. В-третьих, преимуществом технологии CUDA является встроенный механизм блокировок, что обеспечивает простоту реализации данной схемы распараллеливания.

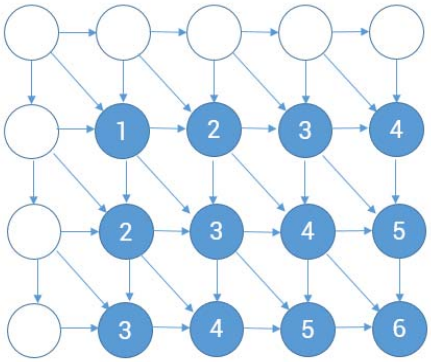


Рисунок 7 Схема работы параллельного алгоритма

Алгоритм для GPU реализован следующим образом. Создается нужное количество блоков CUDA [1] размерностью (`MAX_THREADS, 1, 1`), где `MAX_THREADS` – максимально возможное число потоков для каждого блока. Далее, для каждой строки в отдельном потоке в цикле от `- threadIdx.x + 1` до `n` (где `threadIdx.x` – текущий номер потока, `n` – количество столбцов матрицы) считаем элемент матрицы, если текущий индекс больше или равен 1. Когда очередная итерация цикла завершается, с помощью встроенной фунцкии `__syncthreads()` синхронизируем потоки. Таким образом достигается максимальное использование ресурсов GPU для настоящей задачи и исключается возможность взаимной блокировки процессов.

Настоящая реализация тестировалась на GPU NVIDIA GeForce 740M (2048 Mb memory, 384 cores) и CPU Intel Core i7 3537U (2.0 GHz, 2 cores). В таблице 1 представлены результаты сравнения скорости работы алгоритма для дискретных сигналов разной длины.

Таблица 1 исследование ускорения в зависимости от длины последовательности (время работы – в секундах)

Длина последова-тельности / Реализация	**100**	**200**	**500**	**900**
Python	0.133	0.584	5.327	11.58
C	0.061	0.142	0.351	0.751
CUDA	0.001	0.0025	0.007	0.017
Ускорение (CUDA / C)	61x	56x	50x	44x
Ускорение (CUDA / Python)	131x	237x	492x	670x

5 Заключение

В рамках настоящей работы был разработан быстрый алгоритм совмещения ультразвуковых дефектограмм рельсового пути, основанный на технологии CUDA, позволяющий существенно ускорить процесс работы операторов-дефектоскопистов. Полученные экспериментальные результаты показывают, что достигнут серьезный прирост в производительности, что говорит об эффективности предложенного подхода к распараллеливанию. Целью дальнейшей работы является автоматическое распознавание дефектов в дефектограмме, а также точное определение типа дефекта для дальнейшего анализа специалистов.

Литература

[1] CUDA Zone [Электронный ресурс] // Nvidia Developer: [сайт]. [2016]. URL: https:// developer.nvidia.com/cuda-zone

[2] Eamonn J. Keogh M.J. Derivative Dynamic Time Warping // SIAM International Conference on Data Mining. 2001.

[3] Martens R. C.L. On-line signature verification by dynamic time-warping // IEEE: ICPR. 1996. pp. 38-42.

[4] Wang X., Hui D., Trajcevski G., Scheuermann P., Keogh E. Experimental comparison of representation methods and distance measures for time series data // Data Mining and Knowledge Discovery, 2010. pp. 1-35.

[5] Беллман Р. К.Р. Динамическое программирование и современная теория управления. Москва: Наука, 1969. 118 с.

[6] Маленичев А.А., Сулимова В.В., Красоткина О.В., Моттль В.В., Марков А.А. Применение процедуры парного выравнивания для разметки стыков на ультразвуковой дефектограмме рельсового пути // Известия ТулГУ. Технические науки. 2013. Вып. 9.Ч1. С. 115-127.

[7] Марков А.А., Гараева В.С. Об акустическом контакте в зоне болтовых стыков // Путь и путевое хозяйство, № 12, 2008. С. 15-17.

[8] Марков А.А., Козъяков А.Б., Кузнецова Е.А., Шпагин Д.А. Утраченные и новые технологии контроля рельсов // Путь и путевое хозяйство, № 8, 2013. С. 2-9.

[9] Марков А.А. Ультразвуковая дефектоскопия рельсов. 2-е, переработанное и дополненное-е изд. СПб: Образование - Культура, 2013. 284 с.

[10] Чепрасов Д.Н., Маленичев А.А., Сулимова В.В., Красоткина О.В., Моттль В.В., Марков А.А. Восстановление пропущенных данных на ультразвуковых дефектограммах рельсового пути на основе локально-глобального выравнивания // Машинное обучение и анализ данных, Т. 1, № 12, 2015.

Fast algorithm of combining ultrasonic rail defectograms

Aleksandr Zhukov, Olga Krasotkina,
Anton Malenichev, Valentina Sulimova

The paper is focused on the development of a fast algorithm to combine ultrasonic defectograms of railway track by using multiple channels. An important problem in the analysis of defectograms is to restore the missing data from previous passes. This problem has already been solved, but the algorithm is working slowly. In this paper we developed the fast algorithm for the recovery of missing data. We created a parallel implementation of dynamic programming on the GPU. Analysis of the results showed that the achieved performance is good enough to use algorithm for data processing specialists.

Управление знаниями

Knowledge Management

Управление математическими знаниями: онтологические модели и цифровые технологии

© А.М. Елизаров[1] © А.В. Кириллович[1] © Е.К. Липачев[1] © О.А. Невзорова[2]

[1] Казанский (Приволжский) федеральный университет,
[2] Институт прикладной семиотики АН Республики Татарстан,

Казань
amelizarov@gmail.com alik.kirillovich@gmail.com elipachev@gmail.com
onevzoro@gmail.com

Аннотация

Представлены основные идеи, подходы и уже полученные результаты, развиваемые в рамках проекта по разработке технологий управления математическими знаниями на основе онтологий. Ключевой идеей является разработка специализированных онтологий в области математики, которые составят основу специализированной цифровой экосистемы OntoMath, состоящей из совокупности онтологий, инструментов текстовой аналитики и приложений для управления математическими знаниями.

Выполненные исследования лежат в русле идеологии проекта создания Всемирной цифровой математической библиотеки (World Digital Mathematical Library – WDML). Основное назначение WDML – объединить в распределенной системе взаимосвязанных хранилищ оцифрованные версии всего корпуса математической научной литературы, включая как современные источники, так и источники, ставшие историческими. Настоящая работа развивает названное направление исследований. В ней, в частности, представлена система сервисов автоматической обработки больших коллекций физико-математических документов.

1 Введение

В настоящее время благодаря широкому внедрению информационно-коммуникационных технологий в научно-исследовательскую деятельность при проведении новых исследований стало возможным использование всего корпуса накопленных научных знаний. Последнее предполагает создание комплекса технологий, обеспечивающих оптимальное управление имеющимися знаниями, организацию эффективного доступа к ним, а также совместное и многократное использование новых видов структур знаний.

В области профессиональной математики уже накоплен значительный опыт обработки и использования электронного математического контента в рамках различных проектов создания математических электронных библиотек. Например, в настоящее время одной из крупнейших формальных математических библиотек является Mizar [1], которая представляет собой коллекцию Mizar-статей (документов, подготовленных на формальном языке системы Mizar), содержащих определения, теоремы и доказательства [2, 3]. Следует также отметить важные результаты, связанные с формализацией уровня представлений математических статей. Для этих целей разработан широкий спектр языков представления исходных математических текстов – форматы LaTeX, STeX, XML, специализированные формальные языки, а также программные средства конвертации языков (см., например, [4–6]).

Известны также результаты разработки специализированных математических поисковых систем, например, (uni)quation [7], Springer LaTeX Search [8], Wolfram Formula Search [9].

Названные и многие другие реализованные математические проекты подготовили почву для реализации новой идеи – создания Всемирной цифровой математической библиотеки (World Digital Mathematical Library – WDML; термин введен в 2006 году на Генеральной ассамблее Международного математического союза).

2 Проект WDML

Традиционные библиотеки хранят документы, связывая их библиографическими ссылками, и помогают пользователю найти нужный документ на основе его библиографического описания, ключевых слов и выбранных тематических рубрик. По этой причине традиционные библиотеки обладают существенным ограничением – они не обеспечивают

Труды XVIII Международной конференции DAMDID/RCDL'2016 «Аналитика и управление данными в областях с интенсивным использованием данных», Ершово, 11-14 октября 2016

пользователю непосредственного доступа к элементам математического знания, последний вынужден вручную искать в документах интересующие его математические понятия и выявлять скрытые связи между ними. Появление электронных библиотек сделало работу с документами более эффективной: они предоставляют мгновенный доступ к документам и обеспечивают полнотекстовый поиск по ним. Однако основной принцип работы электронной библиотеки по-прежнему опирается на хранение и поиск документов. Чтобы преодолеть указанное ограничение, проект WDML предлагает новую систему организации и хранения математического знания. В отличие от традиционных электронных математических библиотек, основными элементом этой системы являются не документы, а математические объекты (определения, аксиомы, теоремы, доказательства, уравнения и т. д.), а также логические связи между ними. Информация в WDML должна храниться в формализованном и понятном компьютеру виде, сформированном на основе технологий Семантического Веба. Такой способ управления математическими знаниями позволит создать инструменты для работы непосредственно с объектами математического знания (средства агрегации, семантического поиска, поиска по формулам и идентификации тождественных объектов) [10–13].

В глобальной инициативе WDML можно выделить ряд важнейших направлений, связанных как с организационными усилиями мирового математического сообщества, в том числе издателей математической литературы, так и исследовательских и технологических направлений, нацеленных на разработку и внедрение новых (семантических) технологий представления и обработки математического контента. К числу последних можно отнести:

- агрегирование различных онтологий, индексов, других ресурсов, созданных математическим сообществом, и обеспечение широкого доступа к ним для пополнения и редактирования;
- расширение возможностей доступа к математическим публикациям – не только поиск и просмотр, но и аннотирование, навигация, связывание с другими источниками, организация вычислений, визуализация данных и т. п.

Одной из ключевых идей является разработка классов объектов для адекватного описания и исследования математического содержания (контента). Структурированность математического документа позволяет выделить некоторый набор базовых классов математических объектов (последовательности, функции, преобразования, тождества, символы, формулы, теоремы, утверждения и др.). Как отмечается в проекте WDML, одной из важнейших задач является построение списков математических объектов в разных областях математики.

Переход к представлению внутренней структуры математического знания создает новую парадигму представления, в которой основные акценты смещаются на выделение элементов (классов) и их взаимосвязей, что позволяет создавать различные сетевые концептуальные структуры (например, граф цитирования, граф математических концептов и др.). Выделение классов математических объектов и организация соответствующих репозиториев позволят создать новые вычислительные возможности по обработке данных, такие, как извлечение и обработка формул, поиск близких результатов и т. п.

Семантический поиск позволит по введенному описанию объекта или после выделения термина в статье получить дополнительную информацию (определение, свойства, связи с другими объектами, список документов, в которых встречается объект, указание публикации, где он впервые введен). С помощью такого поиска, например, можно найти все теоремы, в доказательстве которых, прямо или косвенно, используется пятый постулат Евклида. Важно отметить, что в различных документах объект может обозначаться различными терминами.

Инструменты агрегации позволят автоматически собирать объекты, удовлетворяющие заданному критерию и формировать на их основе автоматически пополняемые списки, например, список объектов из заданной предметной области или список теорем, касающихся заданного математического объекта. Эти списки помогут математикам быть в курсе последних достижений, не тратя время на мониторинг всей литературы и не решая повторно уже решенную задачу.

С помощью инструмента поиска по формулам пользователь сможет выделить формулу и получить о ней дополнительную информацию (название, список литературы и т. д.) или ввести уравнение и получить список статей, в которых оно исследуется. При этом формулы в просматриваемых документах могут иметь различное символьное представление.

Инструменты идентификации предназначены для выявления тождественных объектов, которые упоминаются под разными именами и с использованием различной нотации.

Таким образом, основное назначение WDML – объединить в распределенной системе взаимосвязанных хранилищ оцифрованные версии всего корпуса математической научной литературы, включая как современные источники, так и источники, ставшие историческими.

В базовых документах проекта подчеркнуто, что необходимо также обеспечить интеллектуальное извлечение информации для последующей передачи пользователю [12, 13]. В качестве примера назван

сервис, позволяющий пользователю выделить формулу, а затем обратиться к WDML для получения разъяснений и необходимых ссылок.

Для задачи навигации по всему корпусу математических документов ожидаются новые решения, обеспечивающие просмотр и получения дополнительной информации об интересующих объектах: разработка улучшенных механизмов ранжирования документов, в том числе по запросам пользователя; поиск документа, в котором впервые упоминается определенный математический результат, оперативный доступ к справочным ресурсам по теме запроса и т. п.

В целом разработчики проекта WDML полагают, что следующий шаг в развитии и продвижении математики состоит в выходе за пределы традиционных математических публикаций и построении сети информации, основанной на знаниях, содержащихся в этих публикациях.

Одновременно все более востребованными у ученых становятся новые способы обнаружения объектов научного знания непосредственно через Веб, а также инструменты и сервисы, обеспечивающие создание и совместное использование новых видов структур знаний. В контексте концепции связанных данных (Linked Data) и Семантического Веба такие инструменты и сервисы можно использовать для создания «графов сотрудничества» (collaboration graph), которые полезны, например, для вычисления «расстояния сотрудничества» (collaboration distance) между авторами и выделения «близких» документов, что открывает новые возможности тонкой настройки поиска и просмотра (см., например, [14]). Многими авторами (например, [15–17]) подчеркивается важность разработки новых онтологий предметных областей, в частности, математики, поскольку традиционной библиографической каталогизации сегодня уже недостаточно – требуется более глубокая детализация, содержащая описания, созданные с учетом разных точек зрения.

Основные задачи построения WDML и технологии, необходимые для их решения, обсуждены в 2014–2015 гг. широким кругом математиков и закреплены в ряде документов, принятых Всемирным математическим союзом. В частности, одобрено, что следующим шагом в развитии проекта WDML будут выход за пределы традиционных математических публикаций и построение сети информации, основанной на знаниях, содержащихся в этих публикациях. Благодаря сочетанию методов машинного обучения и усилий редакций и редколлегий математических научных журналов, значительная часть информации и знаний (как связанных открытых данных) в глобальном математическом корпусе знаний станет доступной для исследователей через WDML. К реализации названных идей приглашен ряд научно-исследовательских групп по всему миру, в том числе наша группа в Казанском (Приволжском) федеральном университете (КФУ). Представители многих научных групп, задействованных в реализации проекта WDML, приняли участие в симпозиуме, прошедшем в феврале 2016 года в Филдсовском институте (г. Торонто) [18]. Доклад представителей КФУ на этом симпозиуме был посвящен модельным и программным решениям в области семантического представления математического знания [19]. Эти результаты в большой степени коррелируют с общей идеологией проекта WDML в части семантического представления и обработки математических знаний и являются стратегическим направлением исследований Казанской группы, связанным, в частности, с построением экосистемы OntoMath, которая описана ниже.

Отметим, что уровни представления математического знания, соответствующие форматы представления и применяемые семантические модели могут быть описаны, как показано на рис. 1.

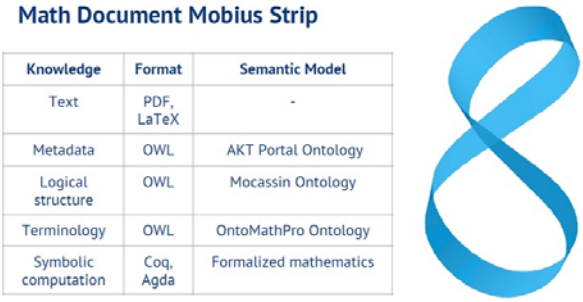


Рисунок 1 Уровни представления математического контента [19]

3 Экосистема OntoMath

OntoMath (http://ontomathpro.org/) – это цифровая экосистема онтологий, инструментов текстовой аналитики и приложений, предназначенная для управления математическими знаниями. В ее состав на текущий момент входят:

- онтология логической структуры математических документов Mocassin;
- онтологии профессиональной математики OntoMathPRO;
- программная платформа для подготовки математического набора связанных данных для публикации в облаке Linked Open Data (LOD);
- сервис семантического поиска по математическим формулам.

Кратко опишем указанные основные элементы.

Семантическое аннотирование математических текстов базируется на онтологии, построенной в рамках проекта Mocassin [20], и онтологии профессиональной математики OntoMath^{PRO} [21].

При разработке последней использовались различные терминологические источники: классические книги, интернет-ресурсы (Wikipedia, журнала «Известия вузов. Математика», а также личный опыт работы профессиональных математиков Казанского университета.

Онтология Mocassin – это онтология логической структуры математических документов. Она разработана В.Д. Соловьевым и Н.Г. Жильцовым и предназначена для автоматического анализа математических публикаций в формате LaTEX [22]. Эта онтология описывает семантику структурных элементов математических документов (например, теоремы, леммы, доказательства, определения и т. д.) и связей между ними. Она разработана с использованием языков OWL2/RDFS [23], которые обеспечили ей выразительные возможности, а также теоретические и практические средства вывода, например, с использованием таких современных машин вывода, как Pellet [24] и FaCT++ [25]. Онтология Mocassin содержит типовые концепты и отношения, эффективно извлекаемые из текстов автоматическими методами [26] (см. рис. 2). Она включает следующие концепты: *DocumentSegment* (Сегмент документа), *Axiom* (Аксиома), *Claim* (Утверждение), *Conjecture* (Гипотеза), *Corollary* (Следствие), *Definition* (Определение), *Equation* (Формула), *Example* (Пример) и др. Математический документ в этой модели рассматривается как набор связанных сегментов, которые являются частью документа, имеют начальную и конечную позиции в тексте, а также характеризуются конкретной функциональной ролью. Онтология структуры научных публикаций по математике описывает семантику сегментов и такие возможные отношения между ними, как *dependsOn* (зависит от), *exemplifies* (является примером для), *hasConsequence* (имеет следствие), *hasSegment* (содержит как сегмент),

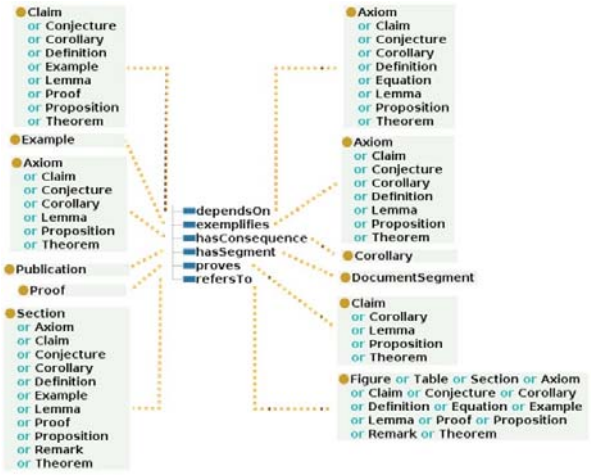


Рисунок 2 Элементы онтологии Mocassin [22]

Онтология профессиональной математики OntoMath^{PRO} – это онтология математического знания, которая организована в виде двух иерархий (см. рис. 3):

- иерархии областей математики: математическая логика, теория множеств, алгебра, геометрия, топология и т. д.;
- иерархии математических объектов: множество, функция, интеграл, элементарное событие, многочлен Лагранжа и т. д.).

OntoMath^{PRO} разработана на языках OWL-DL/RDFS и содержит 3450 классов, 6 типов свойств объектов, 3630 экземпляров свойства IS-A и 1140 экземпляров остальных свойств. Она содержит пять типов отношений: *Класс → Подкласс*, *Определяется с помощью*, *Ассоциативная связь*, *Задача → Метод решения* и *Область математики → Математический объект*. Концепты онтологии содержат их название на русском и английском языках, определение, ссылки на внешние ресурсы из облака Linked Open Data и связи с другими концептами Объектами семантического аннотирования также являются формулы, связанные с формулами фрагменты текста, задающие описания переменных формул [17, 27, 28].

Рисунок 3 Фрагмент онтологии OntoMath^{PRO}

Одним из важных приложений, разработанным на основе указанных выше онтологических моделей, является специальная программная платформа для подготовки математического набора связанных данных для публикации в облаке LOD. Подготовка математического набора связанных данных выполняется на основе разработанных программных инструментов, реализующих комплексный технологический процесс подготовки RDF-набора данных [29]. В качестве экспериментальной коллекции использовались статьи журнала «Известия вузов. Математика» за 1997–2009 гг.

Основными функциями разработанного программного прототипа для публикации данных в облаке LOD являются:

- индексирование математических статей в формате LaTEX в виде LOD-совместимых RDF-данных;
- извлечение метаданных статьи в виде концептов онтологии AKT Ontology [30];
- извлечение логической структуры документа с использованием онтологии Mocassin;
- извлечение экземпляров математических сущностей в виде концептов онтологии OntoMathPRO и связывание с ресурсами DBPedia;
- распознавание семантики формул через связывание полученных экземпляров математических сущностей с математическими выражениями и формулами в тексте;
- установление взаимосвязи между опубликованными RDF-данными и существующими наборами данных LOD.

Разработанная технология имеет следующие отличительные особенности:

- математический RDF-набор строится на основе коллекции математических статей на русском языке;
- построенный RDF-набор помимо метаданных статей включает специальные семантические знания: знания, формируемые в результате специальной обработки математических формул – семантического связывания текстовых определений переменных формул с их символьными обозначениями; знания, связанные с идентификацией в тексте экземпляров онтологии OntoMathPRO; также знания о структурных элементах математической статьи.

Архитектура прототипа программной системы включает 8 модулей, которые могут быть сгруппированы в следующие подсистемы:

- преобразование формата;
- аннотирование текста;
- семантическое аннотирование;
- аннотирование метаданных;
- генерация RDF;
- связывание.

Подробное описание названных модулей дано в [29].

Другим важным инструментом является разработанный нами семантический поисковик по математическим формулам [27]. Отметим, что известные сегодня сервисы поиска по формулам (такие, как (uni)quation, Springer LaT_EX Search, Wikipedia Formula Search, Wolfram Formula Search) являются синтаксическими и позволяют лишь найти формулы, содержащие заданные фрагменты (например, $(a+b)^2$). В отличие от них семантический поисковый сервис системы OntoMath решает задачу поиска формул по именам ее символьных переменных, в частности, выполняет поиск формулы, содержащей переменную, обозначающую заданное математическое понятие (например, формулы, содержащие обозначение угла или связь давления и массы).

Этот сервис использует семантическое представление документа, построенное с помощью платформы семантической публикации, описанной выше. Названное представление отражает связи между элементами логической структуры документа, используемой терминологией, переменными и формулами. При этом пользователь может ограничивать контекст поиска, например, искать только в текстах определений и формулировках теорем.

Еще одним приложением экосистемы OntoMath является рекомендательная система для коллекций физико-математических документов, которая для каждого из них строит список «близких» документов (см. [31]). Традиционно список «близких» документов формируется на основе выбранной меры близости ключевых слов, приведенных авторами, а также библиографических ссылок, имеющихся в документах. Этот подход имеет ряд недостатков:

- список ключевых слов может быть неполным или отсутствовать;
- проблема омонимии: одно и то же понятие может обозначаться разными ключевыми словами, например, «полином» и «многочлен»;
- не учитываются родовидовые отношения между понятиями, например, статья с ключевым словом «пятиугольник» не будет определена как «близкая» к статье с ключевым словом «многоугольник»;
- существенна привязка к языку, например, статья с ключевым словом «матрица» не будет «близка» к статье с ключевым словом «matrix».

Таким образом, авторского списка ключевых слов недостаточно, и необходим более глубокий анализ содержания документа. Одним из методов такого анализа является терминологическое аннотирование, основанное на онтологиях предметных областей (например, [32]).

В разработанной рекомендательной системе реализуется следующие основные этапы:

- на основе онтологии $\text{OntoMath}^{\text{PRO}}$ производится извлечение ключевых слов из документов коллекции;
- каждая публикации представляются в виде вектора, компоненты которого соответствуют концептам онтологии;
- значение компоненты – это вес соответствующего понятия в данной статье (вычисляется с использованием количества его упоминаний в тексте статьи и количества упоминаний связанных понятий) (см. [31]);
- в качестве меры близости между публикациями используется косинусная мера близости между их векторами.

Заключение

Описаны основные идеи, подходы и полученные результаты по разработке технологий управления математическими знаниями на основе специализированных онтологий в области математики. Эти решения составляют основу специализированной цифровой экосистемы OntoMath, которая состоит из совокупности онтологий, инструментов текстовой аналитики и приложений для управления математическими знаниями. Выполненные исследования лежат в русле проекта создания Всемирной цифровой математической библиотеки World Digital Mathematical Library.

Благодарности

Работа выполнена при финансовой поддержке РФФИ (проекты №№ 15-07-08522, 15-47-02472).

Литература

[1] http://www.mizar.org/.

[2] A. Naumowicz, A. Kornilowicz. A brief overview of Mizar. In: S. Berghofer et al. (Eds.), TPHOLs 2009, Lecture Notes in Computer Science 5674, Springer-Verlag Berlin Heidelberg, p. 67–72, 2009.

[3] G. Bancerek, C. Bylinski, A. Grabowski, A. Kornilowicz, R. Matuszewski, A. Naumowicz, K. Pak, J. Urban. Mizar: State-of-the-Art and Beyond. In M. Kerber et al. (Eds.), Intelligent Computer Mathematics, CICM 2015, Lecture Notes in Artificial Intelligence 9150, p. 261–279, 2015.

[4] M. Kohlhase. Using LaTeX as a Semantic markup format. Mathematics in Computer Science (2:2); p. 279–304, Birkhäuser 2008.

[5] H. Stamerjohanns, D. Ginev, C. David, D. Misev, V. Zamdzhiev, M. Kohlhase. Conversion d'articles en LaTeX vers XML avec MathML: une étude comparative. Cahiers GUTenberg, 51; p. 7–28, 2010, http://cahiers.gutenberg.eu.org/cg-bin/article/CG_2008_51_7_0.pdf.

[6] M. Iancu, M. Kohlhase, F. Rabe, J. Urban. The Mizar Mathematical Library in OMDoc: Translation and Applications; Journal of Automated Reasoning, 50:2; p. 191–202, Springer Verlag 2013.

[7] Сайт (uni)quationalpha math expression search engine. http://uniquation.com/en/.

[8] Сайт LaTeX Search Beta. http://latexsearch. com/.

[9] The Wolfram Functions Site. http://functions. wolfram.com/.

[10] Digital Mathematics Library: a vision for the future. International Mathematical Union, 2006. http://www.mathunion.org/fileadmin/IMU/Report/dml_vision.pdf.

[11] P. J. Olver. What's happening with the World Digital Mathematics Library? http://www.math.umn. edu/~olver/t_/wdmlb.pdf.

[12] Developing a 21st century global library for mathematics research. Washington, D.C.: The National Academies Press, 2014. 131 p. arxiv.org/pdf/1404.1905; http://www.nap.edu/catalog/18619/developing-a-21st-century-global-library-for-mathematics-research.

[13] P. J. Olver. 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, V. 1, p. 773–785, 2014.

[14] R. Todeschini, A. Baccini. Handbook of bibliometric indicators: quantitative tools for studying and evaluating research. Wiley-VCH Verlag, 2016.

[15] S. Staab, R. Studer (Eds.). Handbook on ontologies. Berlin, Heidelberg: Springer Verlag, 2003, 2009. 811 p.

[16] C. Lange. Ontologies and languages for representing mathematical knowledge on the Semantic Web. Semantic Web Journal, 2010. http://www.semantic-web-journal.net/content/ontologies-and-languages-representing-mathematical-knowledge-semantic-web.

[17] A. Elizarov, A. Kirillovich, E. Lipachev, O. Nevzorova, V. Solovyev, N. Zhiltsov. Mathematical knowledge representation: semantic models and formalisms. Lobachevskii Journal of Mathematics, V. 35, No 4, p. 347–353, 2014.

[18] Semantic representation of mathematical knowledge workshop, 5 February 2016: https://www.fields.utoronto.ca/programs/scientific/15-16/semantic/.

[19] A. M. Elizarov, N. G. Zhiltsov, A. V. Kiril-lovich, E. K. Lipachev, O. A. Nevzorova, V. D. So-lovyev. 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.

[20] Mathematical Semantic Search engINe (MocaSSIN). https://code.google.com/archive/p/mocassin/.

[21] OntoMathPRO: A Hub for Math Linking Open Data (LOD). https://github.com/CLLKazan/OntoMathPro.

[22] V. Solovyev, N. Zhiltsov. Logical structure analysis of scientific publications in mathematics. In: Proceedings of the Int. Conf. on Web Intelligence, Mining and Semantics (WIMS'11). ACM, p. 21:1–21:9, 2011. http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.473.9071&rep=rep1&type=pdf.

[23] OWL 2 Web Ontology Language. RDF-Based Semantics (Second Edition). W3C Recommendation 11 December 2012. https://www.w3.org/2012/pdf/REC-owl2-rdf-based-semantics-20121211.pdf.

[24] Pellet: An Open Source OWL DL reasoner for Java. http://clarkparsia.com/pellet.

[25] FaCT++. http://owl.man.ac.uk/ factplusplus/.

[26] О. А. Невзорова, Е. В. Биряльцев, Н. Г. Жильцов. Коллекции математических

текстов: аннотирование и применение в поисковых задачах. Искусственный интеллект и принятие решений, № 3, с. 51–62, 2012.
[27] O. Nevzorova, N. Zhiltsov, A. Kirillovich, E. Lipachev. OntoMathPro ontology: a linked data hub for mathematics. Communications in Computer and Information Science, V. 468, p. 105–119, 2014.
[28] А. М. Елизаров, Е. К. Липачёв, О. А. Невзо-рова, В. Д. Соловьев. Методы и средства семантического структурирования электронных математических документов. Докл. РАН, Т. 457 (6), с. 642–645, 2014.
[29] O. Nevzorova, N. Zhiltsov, D. Zaikin, O. Zhib-rik, A. Kirillovich, V. Nevzorov, E. Birialtsev. Bringing Math to LOD: a semantic publishing platform prototype for scientific collections in mathematics. 12th Int. Semantic Web Conference, Sydney, NSW, Australia, October 21–25, 2013, Proceedings, Part I. Lecture Notes in Computer Science, Vol. 8218. Springer Berlin Heidelberg, p. 379–394, 2013.
[30] AKT Ontology. http://dream.inf.ed.ac.uk /projects/dor/akt/akt.html.
[31] А. М. Елизаров, А. Б. Жижченко, Н. Г. Жильцов, А. В. Кириллович, Е. К. Липачёв. Онтологии математического знания и рекомендательная система для коллекций физико-математических документов. Докл. РАН, Т. 467, № 4, с. 392–395, 2016.
[32] А. М. Елизаров, Н. Г. Жильцов, А. В. Кириллович, Е. К. Липачёв. Терминологическое аннотирование и рекомендательный сервис в системе управления физико-математическим контентом. Труды XVII Межд. конф. DAMDID/RCDL’2015 «Аналитика и управление данными в областях с интенсивным использованием данных». Обнинск: ИАТЭ НИЯУ МИФИ, с. 347–350, 2015.

Mathematical knowledge management: ontological models and digital technology

Alexander M. Elizarov, Alexander V. Kirilovich, Evgeny K. Lipachev, Olga A. Nevzorova

This paper is discussed basic ideas, approaches and the results obtained in the research project the objective of which is to develop mathematical knowledge management technologies based on ontologies. We are developing the digital ecosystem OntoMath for mathematical knowledge management, which includes a set of specialized ontologies, text analytics tools and applications for managing mathematical knowledge. The results obtained are close to main problems declared in the World Digital Math Library (WDML) project. The main purpose of WDML is to build a global system of linked repositories for saving all digital mathematical documents, including contemporary and historic sources. This paper is devoted to decisions of some problems in this global initiative. In particular, we developed the program services for processing large collections of mathematical papers.

Визуальная аналитика многомерных динамических данных

© Д.Д. Попов[1] © И.Е. Мильман[1] © В.В. Пилюгин[1] © А.А. Пасько[2]

[1]Национальный исследовательский ядерный университет «МИФИ»,

Москва, Россия

[2]Британский национальный центр компьютерной анимации при университете Борнмута,

Борнмут, Великобритания

dpopovmephi@gmail.com igalush@gmail.com VVPilyugin@mephi.ru

apasko@bournemouth.ac.uk

Аннотация

В статье рассматривается задача анализа изменения некоторых объектов. Каждый объект характеризуется набором из n численных параметров. Инструментом проведения анализа является метод визуализации. Получение интересующих аналитика суждений о рассматриваемых объектах осуществляется за несколько шагов. На первом шаге строится геометрическая интерпретация исходных данных. Затем построенная математическая модель подвергается ряду преобразований. Эти преобразования соответствуют решению первой задачи метода визуализации – непосредственно визуализации исходных данных. Далее человеку предлагается провести анализ полученных изображений и интерпретировать результаты по отношению к рассматриваемым объектам. Иными словами, решение задачи анализа осуществляется с помощью пространственного моделирования исходных данных и последующего исследования аналитиком построенных пространственных объектов.

Предложен алгоритм решения задачи. Описана разработанная интерактивная прикладная программа визуализации, реализующая этот алгоритм. Продемонстрировано, как в процессе работы с программой пользователь может построить суждения об образовании и разрушении кластеров и сгустков из объектов, а также найти инварианты в изменении исходных данных.

1 Введение

Среди различных методов анализа данных в настоящее время важное место занимают перспективные методы визуальной аналитики. Под визуальной аналитикой понимается решение задач анализа данных с использованием способствующего интерактивного визуального интерфейса. Этот термин ввёл Джеймс Томас (James Thomas) в [3], чем обратил внимание научной общественности на роль визуального представления данных в решении исследовательских задач.

Одной из форм визуальной аналитики является решение задач анализа данных методом визуализации. Теоретические обобщения этого метода на примере научных данных рассмотрены в работе [7]. Этот метод заключается в последовательном, в общем случае многократном, решении двух задач:

1. задачи визуализации исходных данных;
2. задачи анализа полученных графических изображений с последующей формулировкой суждений относительно исходных данных.

В данной работе рассмотрен ряд вопросов теории и практики дальнейшего развития этого метода применительно к анализу многомерных динамических данных. Здесь многомерными динамическими данными мы называем числовые данные, представляемые в табличном виде. Они описывают некоторые изменения состояния рассматриваемой совокупности объектов, происходящие с течением времени.

2 Подходы к визуализации многомерных данных

За последние 30 лет появилось большое количество научных статей и книг, посвящённых подходам к визуализации. Преимущественно данную тему освещают зарубежные авторы. Найдены публикации европейских и американских исследователей, которые датируются 70-ми, 80-ми годами XX века. Отечественные учёные активно подключились к изучению данного вопроса

Труды XVIII Международной конференции DAMDID/RCDL'2016 «Аналитика и управление данными в областях с интенсивным использованием данных», Ершово, 11-14 октября 2016

значительно позже (основной информационный массив российского авторства относится к 2000-2005 гг.).

Среди работ об анализе многомерных динамических данных можно обнаружить большое количество статей посвящённых динамике наноструктур, как, например, [1] и [2]. В [5] проводится визуальный анализ свойств пространственно-временных структур. Данные об объектах, указанных в этих статьях, имеют трёхмерную геометрическую интерпретацию. Также в [5] констатируется существование проблемы визуализации данных, имеющих размерность более 4[1], адекватной восприятию пользователя. Нас будут интересовать пространства размерности больше 3.

В [8] предложен способ визуализации многомерных данных о состоянии пациента. Производится анализ *одного* объекта, характеризующегося множеством (мощность этого множества > 3) числовых параметров, значения которых изменяются с течением времени. Метод, описываемый в статье, направлен на решение задачи анализа изменений *множества* многомерных объектов. При чём наибольший практический интерес представляют множества большой размерности.

В работе [4] специфика визуализируемых многомерных динамических данных заключается в следующем: 2 или 3 числовые компоненты этих данных имеют пространственную природу. В русскоязычных трудах также рассматриваются способы визуального анализа данных, имеющих такую особенность. Рассматриваемые в них программные комплексы называются геоинформационными системами (ГИС). В ГИС каждому объекту ставится в соответствие некоторый геометрический образ, изображающийся на плоской или объёмной карте местности. В качестве таких геометрических образов могут выступать круги, квадраты, шары, кубы, линии, показывающие траектории движения исследуемых объектов. Числовые компоненты данных, не пространственной природы представляются в отдельных окнах визуального пользовательского интерфейса либо поочерёдно, в соответствии с запросом пользователя, отражаются на карте местности за счёт изменения оптических параметров геометрических объектов (например, цвет, прозрачность) или их размеров, формы. В данной статье числовые характеристики объектов имеют произвольную природу.

Указанные выше программы и методы анализа многомерных динамических данных, предполагающие их использование, различаются не только моделями исходных данных и подходами к их визуализации. Упомянутые различия обусловлены поставленными задачами анализа этих данных. Нас будет интересовать исследование данных с целью выявления схожих объектов среди рассматриваемой совокупности. Изучается, как изменяются множества подобных объектов, а также, что остаётся неизменным с течением времени.

3 Постановка задачи

Пусть имеется множество из *m* объектов, каждый из которых характеризуется *n* параметрами.

Мы предполагаем, что в рамках рассматриваемой задачи исходно задано *k* таблиц вида:

Таблица 1 *j*-ая таблица с данными об объектах

	Параметр 1	Параметр 2	…	Параметр n
Объект 1	x_{11}^j	x_{12}^j	…	x_{1n}^j
Объект 2	x_{21}^j	x_{22}^j	…	x_{2n}^j
…	…	…	…	…
Объект m	x_{m2}^j	x_{m2}^j	…	x_{mn}^j

Каждая таблица содержит значения параметров объектов в некоторый момент времени. То есть *j*-ая таблица содержит данные для момента времени t_j. При этом будем считать, что $t_1 < t_2 < \cdots < t_k$. x_{il}^j – значение *l*-ого параметра *i*-ого объекта в соответствующий *j*-ой таблице момент времени t_j ($l = \overline{1,n}, i = \overline{1,m}, j = \overline{1,k}$).

3.1 Геометризация данных

Для решения задачи анализа перейдём к её геометрической интерпретации. Будем считать, что исходные данные являются точками многомерного евклидова пространства E_n с заданной метрикой $\rho(x, y)$. При этом:

- Строка таблицы интерпретируется как точка многомерного евклидова пространства.
- Столбцы таблиц интерпретируются как координаты точек многомерного евклидова пространства.
- Расстояние в евклидовом пространстве трактуется как мера различия объектов.

Используя введённое в пространстве расстояние, можем выделить подмножества точек: сгустки и кластеры. Определения этих подмножеств даны в [6]. Координаты точек известны в моменты времени $t_1, t_2, \dots, t_k$.

Введём понятие *геометрического процесса*. *Геометрический процесс* – это множество точек пространства, координаты которых зависят от времени. Таким образом, изначально имеем дискретный геометрический процесс $P = P(t_j)$. $P(t_j)$ задан на множестве $T = \{t_1, t_2, \dots, t_k\}$.

$$P(t_j) = \{p_1(t_j), p_2(t_j), \dots, p_m(t_j)\},$$

где $p_i(t_j)$ - n-мерная точка.

$$p_i(t_j) = \left(p_i^1(t_j), p_i^2(t_j), \dots, p_i^n(t_j)\right),$$

[1] В случае данных, имеющих размерность 4 в качестве графической интерпретации выступают изоповерхности, то есть 3 числовые характеристики задают координаты точек привычного для человека трёхмерного пространства, и точки с одинаковым значением 4-го параметра выделяют одним цветом.

Рисунок 1 Метод визуализации

где $p_i^j(t_j) = x_{il}^j$ – l-ая координата i-ой точки.

Для каждой пары точек определено расстояние

$$c\left(p_{i_1}(t_j), p_{i_2}(t_j)\right) \geq 0,\ i_1, i_2 = \overline{1,m},\ i_1 \neq i_2$$

$$\rho(p_i, p_j) \geq 0$$

Стоит отметить, что расстояние, то есть метрику евклидова пространства, целесообразно выбирать исходя из известных особенностей исходных данных. Существует множество исследовательских работ, описывающих ситуацию, когда в процессе изучения некоторого набора данных эмпирически подбиралась метрика, которая наилучшим образом справляется с поставленной задачей анализа, например с задачей кластеризации. В разработанной программе, подробно о которой будет сказано позже, имеется возможность использовать евклидову метрику, расстояние Минковского, расстояние Чебышева, в частности, и расстояние Махаланобиса.

4 Преобразования геометрических процессов

Для решения поставленной задачи возникает необходимость в преобразованиях исходного дискретного геометрического процесса.

$P(t_j)$ представляет собой упорядоченный по времени набор описаний известных состояний множества исследуемых объектов. Сами же объекты существуют и изменяются непрерывно. Исходя из этого факта, необходимо иметь инструмент преобразования дискретного процесса в непрерывный. Таким инструментом служит интерполяция.

4.1 Интерполяция

Задача интерполяции состоит в поиске такой функции F из заданного класса функций, что: $F(t_j) = P(t_j)$, где $t_j \in T$.

Так как

$$P(t_j) = \{p_1(t_j), p_2(t_j), \dots, p_m(t_j)\},$$

$$p_i(t_j) = \left(p_i^1(t_j), p_i^2(t_j), \dots, p_i^n(t_j)\right),$$

то поиск F заключается в поиске таких f_i^i, принадлежащих заданному классу, что $f_i^i(t_j) = p_i^i(t_j)$.

В случае требования непрерывности f_i^i в результате интерполяции исходного дискретного геометрического процесса получаем непрерывный геометрический процесс $P(t)$.

Пусть f_i^i является алгебраическим двучленом, тогда при любом $t \in [t_h, t_{h+1}]$, $1 \leq h \leq k-1$, f_i^1 будет рассчитываться по формуле:

$$f_i^i(t) = p_i^i(t_h) + \frac{p_i^i(t_{h+1}) - p_i^i(t_h)}{t_{h+1} - t_h}(t - t_h).$$

Интерполяция алгебраическим двучленом или кусочно-линейная интерполяция является самым простым видом интерполяции для которого выполняется свойство непрерывности.

За неимением априорной информации об изменении исходных данных с течением времени выбор способа интерполяции делается в сторону кусочно-линейной, так как она не требует сложных вычислений и при достаточно малых интервалах времени между имеющимися данными хорошо описывает наблюдающуюся зависимость.

4.2 Дискретизация непрерывных геометрических процессов

Пусть $P(t)$ – непрерывный геометрический процесс, тогда $P(t_0)$ – его временное сечение. t_0 принадлежит области определения $P(t)$.

Из $P(t)$ можно построить дискретный геометрический процесс, заданный в k' моментах времени. Для этого выберем моменты времени $\Phi_1, \Phi_2, \dots, \Phi_{k'}$ и определим $P'(\Phi_h)$ как совокупность временных сечений $P(t)$: $P'(\Phi_h) := \{P(\Phi_h) | \Phi_h \in \{\Phi_1, \Phi_2, \dots, \Phi_{k'}\}\}$.

5 Метод визуализации в рамках решаемой задачи

Метод визуализации подразумевает последовательное решение двух задач, представленных на рисунке 1.

Для описания первой задачи вводится важнейшее понятие метода – *пространственная сцена*. Пространственная сцена – это совокупность пространственных объектов с заданными геометрическими и оптическими параметрами. Описание геометрических параметров объектов сцены, например, их форма, размеры,

пространственное расположение называется *геометрической моделью сцены*. *Оптической моделью сцены* называется описание оптических параметров, например, цвет, прозрачность объектов.

Решение первой задачи подразумевает построение отображения исходных данных на пространственную сцену. Успех проведения анализа напрямую зависит от того насколько удачным окажется построенное отображение. На этом этапе необходимо выбрать такие модели, которые бы способствовали выявлению закономерностей в данных о рассматриваемых объектах. С одной стороны они должны соответствовать целям визуального анализа, например, анализа формы или взаимного расположения, а с другой обеспечивать возможность последующей интерпретации результатов визуального анализа по отношению к рассматриваемым объектам.

В рамках рассматриваемой задачи задание параметров сцены, атрибутов визуализации и получение изображений или анимационных фильмов происходит до тех пор, пока аналитик не рассмотрит достаточное их количество для формирования некоторого суждения об изменении взаимного расположения заданного множества точек с течением времени. Атрибуты

Рассмотрим подробнее процесс визуализации исходных данных.

5.1 Задание динамических исходных данных

Задаётся дискретный геометрический процесс $P(t_j)$, для которого будет проводиться визуализация.

5.2 Фильтрация

На этом шаге исходные данные проходят предварительную обработку.

Для решения поставленной задачи исходный дискретный процесс $P(t_j)$ линейно интерполируется, как описано в 4.1. В результате фильтрации получается непрерывный процесс $P(t)$.

5.3 Мэппинг

На этом этапе исходным данным ставится в соответствие динамическая пространственная сцена $S(t) =< G(t), O(t) >$, где $G(t)$ – описание геометрии сцены, а $O(t)$ – описание оптических параметров сцены.

В каждый момент времени t, $S(t)$ соответствует $P(t), t \in [t_1, t_k]$.

$$G(t) = \{Sph_1(t), \dots, Sph_m(t), Cyl_1(t), \dots, Cyl_e(t)\},$$

где $Sph_i(t)$ – сфера, соответствующая точке $p_i(t)$; $Cyl_j(t)$ – цилиндр, соединяющий 2 сферы. Цилиндры соединяют только такие сферы $Sph_{i_1}(t), Sph_{i_2}(t)$, для соответствующих точек которых выполняется $c\left(p_{i_1}(t_j), p_{i_2}(t_j)\right) \le d$, d – некоторая константа, выбор которой в процессе решения задачи осуществляется аналитиком.

$$O(t) = \{SphC_1(t), \dots, SphC_m(t), CylC_1(t), \dots, CylC_e(t)\},$$

где $SphC_1(t) = \dots = SphC_m(t) = SphC$ – цвета сфер, $SphC$ так же в процессе решения задачи задаётся аналитиком, $CylC_1(t), \dots, CylC_e(t)$ – цвета цилиндров.

Значение цвета в кодировке RGB рассчитывается по формулам:

$$R(t) = 255\left(1 - \frac{c(p_{i_1}(t), p_{i_2}(t))}{d}\right),$$

$$G(t) = 150\frac{c(p_{i_1}(t), p_{i_2}(t))}{d},$$

$$B(t) = 255\frac{c(p_{i_1}(t), p_{i_2}(t))}{d}.$$

5.4 Рендеринг

Понятие пришло из дисциплины машинная графика. Результатом рендеринга является проекционное изображение I сцены S.

Каждому t соответствует пространственная сцена, а значит и проекционное изображение:

$$I(t) = I(S(t), A(t)),$$

$A(t)$ – атрибуты визуализации: камера, освещение, размер получаемого изображения и другие.

6 Описание алгоритма решения задачи

Алгоритм решения задачи состоит из следующих этапов:

1. ввод динамических исходных данных;
2. их интерполяция и получение непрерывного геометрического процесса;
3. задание d и статических параметров сцены: радиус сфер, радиус цилиндров, подпространство для проекции, цвет сфер;
4. построение непрерывного пространственного процесса, соответствующего непрерывному геометрическому процессу, с использованием заданных статических параметров;
5. построение временных сечений непрерывного пространственного процесса, то есть его дискретизация;
6. задание камеры и других атрибутов визуализации для рендеринга;
7. построение проекционных изображений дискретного пространственного процесса;
8. дальнейшее сохранение проекционных изображений в галерею сечений или создание анимационного фильма;

В случае, если аналитик не может сделать интересующего его суждения относительно исходных данных по полученным проекционным изображениям, алгоритм предусматривает возвраты на этапы 3 и 6.

Вышеописанный алгоритм представлен на рисунке 2.

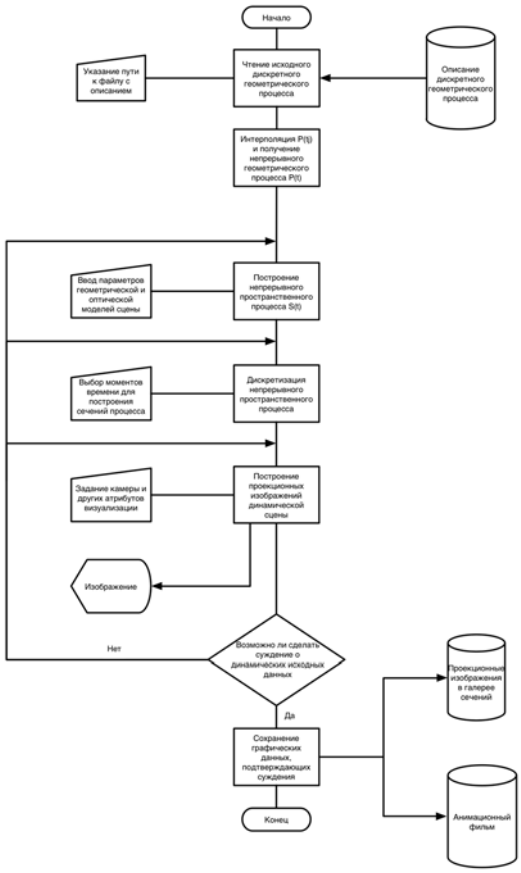


Рисунок 2 Алгоритм решения задачи

Следует отметить, что важным для практики случаем является анализ статических многомерных данных. Нетрудно видеть, что анализ статических многомерных данных можно рассматривать как частный случай анализа динамических многомерных данных. Он соответствует исходному дискретному геометрическому процессу, заданному в некоторый (единственный) момент времени, то есть $T = \{t_1\}$.

Алгоритм предусматривает взаимодействие аналитика и компьютера. Предлагается следующее распределение функций между ними:

- компьютер в процессе решения задачи анализа осуществляет функции расчёта расстояний, построения проекций и графиков зависимости координат от времени;
- аналитик осуществляет зрительное восприятие проекционного изображения на мониторе, анализирует взаимное расположение многомерных точек, выделяет подмножества и задаёт параметры визуализации.

Для реализации данного алгоритма была разработана соответствующая интерактивная программа. Она реализована на базе программного продукта Autodesk 3ds Max, с использованием внутреннего интерпретируемого языка MAXScript. Так же использовалась библиотека, написанная на языке C# в Visual studio 2013.

Эта программа позволяет аналитику совершать следующие действия:

- Ввод исходных данных (описание дискретного геометрического процесса).

  Осуществляется один раз при начале работы с программой.
- Кусочно-линейная интерполяция дискретного геометрического процесса.
- Построение динамической сцены, соответствующей непрерывному геометрическому процессу, и её анимационная визуализация.

  Графическое проецирование выполняется в окне 3ds Max с помощью стандартного рендерера.
- Получение информации о точках.

  Информацию о точках можно получать в виде таблицы, строки которой соответствуют каждой из отображаемых точек, а столбцы — их координатам. При этом, строки таблицы закрашены тем цветом, которым закрашены точки, т.е. в зависимости от цвета можно определять к какому подмножеству относится данная точка. Эти таблицы можно получать для каждого рассматриваемого момента времени.
- Измерение максимального внутрикластерного расстояния между двумя точками (в исходном n-мерном пространстве).
- Задание параметров сцены.

  Параметрами сцены являются: радиус сфер, поставленных в соответствие исходным точкам, их цвет, радиус цилиндров и трёхмерное проекционное пространство. Указанные параметры задаются в начале работы программы, их можно изменять в процессе анализа.
- Использование стандартных средств 3ds Max для работы с пространственной сценой. Так, например, есть возможность использовать следующие средства:

  а) аффинные преобразования сцены;

  б) наложение разнообразных фильтров на изображение;

  в) преобразование оптических характеристик сцены.
- Объединение точек в подмножество.

  Каждому подмножеству даётся цвет, который в дальнейшем будет обозначать данное подмножество во время работы программы.
- Построение и сохранение полученных графических проекций многомерных точек.

  При анализе удалённых точек важным является то, какие именно координаты вносят больший вклад в расстояние — происходит

ли это за счёт всех координат или за счёт большого отличия только нескольких координат. Для определения этого, предлагается строить графические проекции исходного множества (p_i^l, p_i^l) и, меняя l, просматривать все такие проекции.

- Создание анимационного фильма.
- Построение графиков зависимости координат точек от времени.

При проецировании многомерных точек на трёхмерное пространство, особенно в случае большой размерности, проекции точек могут оказаться очень близко друг к другу. Для того, чтобы различить близкие сферы на получаемых изображениях аналитику предлагается воспользоваться такими возможностями изменения пространственной сцены как задание радиусов сфер и цилиндров. Уменьшив их, пользователь программы может более детально изучить точки, проекции которых оказались на малых расстояниях. Кроме того, в случае, если две различные многомерные точки проецируются в одну точку трёхмерного пространства, программа в процессе построения пространственной сцены немного сместит центры соответствующих сфер друг относительно друга и раскрасит их в цвет, отличный от цвета выбранного для сфер сцены по умолчанию.

7 Пример использования разработанной программы

С помощью созданной программы был произведён анализ 81 кредитной организации. Данные представляли собой ежемесячные отчёты этих организаций по 9 показателям за 13 месяцев 2013-2014 гг. Данные за первый отчётный месяц соответствуют $t = 1$, за последний – $t = 13$. Исследовалась принадлежность кредитных организаций в их геометрической интерпретации к сгусткам и кластерам, что позволяет судить об их схожести.

На рисунке приведён ряд изображений, позволяющих сделать вывод, что за промежуток времени $t \in [7,8]$ удалённая точка, соответствующая выделенной белой окружностью сфере, присоединилась к сгустку.

Как отмечается в [6], рассматриваемая в примере задача, актуальна при проведении аудита кредитных организаций. Априорно аналитик знает «хорошие» организации или же выделяет их в процессе анализа. Например, в первом приближении за множество «хороших» можно принять те, соответствующие точки которых принадлежат одному кластеру. Далее исследуются изменения, происходящие с этим множеством. Например, какая-то точка покинула или присоединилась к выделенному кластеру. Организация, соответствующая этой точки попадает под особое внимание аналитика, сделанное им наблюдение может послужить причиной более детального изучения деятельности этой организации. Также для исследователя важны изменения, происходящие внутри одного сгустка или кластера. Если в нём присутствуют точки, для которых с течением времени близость к другим точкам кластера резко изменяется, т.е. они быстро стремятся покинуть кластер или наоборот приближаются к другим точкам, то соответствующие организации также попадают под пристальный контроль. В разработанной программе аналитик фиксирует эти точки, наблюдая за изменением цвета цилиндров, соединяющих сферы.

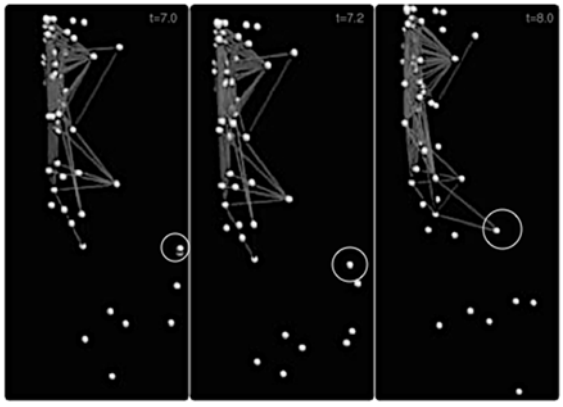


Рисунок 3 Кадры полученного анимационного фильма

Было замечено, что для одной из построенных моделей сцен наблюдается следующая закономерность. Большинство сфер, за исключением выделенных белыми окружностями на рисунке, лежат в одной плоскости.

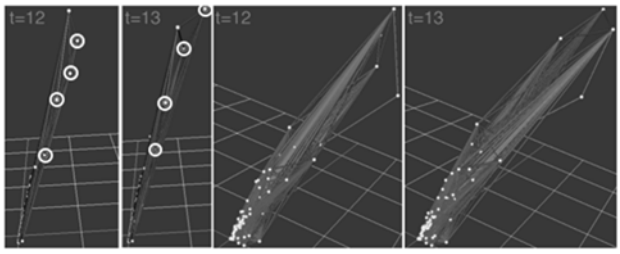


Рисунок 4 Проекционные изображения пространственной сцены

Это позволило выделить зависимость между тремя финансовыми показателями организаций:

X – прощёная задолженность в кредитном портфеле,

Y – выпущенные облигации,

Z – активы нетто.

Она описывается уравнением плоскости:

$$2{,}231X\text{-}72{,}672Y + 20{,}3624Z - 2{,}58513 = 0.$$

Выделенные белыми окружностями сферы и соответствующие им кредитные организации этой зависимости не подчиняются. Зависимость имеет место при $t \in [1,9] \cup [12,13]$.

Заключение

Итак, в рамках данной работы:

- Были созданы математические модели исходных данных, операций над ними.
- Разработан алгоритм решения задачи, на его основе написана программа.
- На примере экономических показателей кредитных организаций продемонстрирован анализ изменения подобия, а так же возможность поиска инвариантов геометрических процессов с помощью разработанной программы.

Инвариантами здесь называются такие свойства исследуемых данных, которые остаются неизменными либо на всём промежутке времени, либо на некоторой его части. В приведённом примере неизменной на выделенных отрезках времени остаётся зависимость трёх координат – показателей финансовых организаций.

В качестве дальнейшего развития функциональных возможностей рассматриваются различные способы интерполяции дискретных геометрических процессов, разработка пользовательского инструментария, позволяющего производить дополнительные построения в пространственной сцене. Например, предоставление возможности построения плоскостей, сфер и других геометрических примитивов.

В настоящей работе не представлены классические методы анализа многомерных данных, предполагающие численное моделирование и автоматический интеллектуальный анализ без участия человека. Тем не менее в настоящее время авторами изучаются эти подходы с целью последующего развития системы до уровня «численно-визуальной» системы анализа данных.

Литература

[1] Grottel S. Visual Verification and Analysis of Cluster Detection for Molecular Dynamics / S. Grottel [et al.]. // IEEE Transactions on Visualization and Computer Graphics. – 2007. – Vol. 13. – №6. – P. 1624-1631.

[2] Sourina O., Korolev N. Visual Mining and Spatio-Temporal Querying in Molecular Dynamics / O. Sourina, N. Korolev // Journal of All rights reserved Computational and Theoretical Nanoscience. – 2005. – Vol. 2. – P. 1-7.

[3] Thomas J., Cook K. Illuminating the Path: Research and Development Agenda for Visual Analytics. – IEEE-Press, 2005. – 185 p.

[4] Wallner G. PLATO: A visual analytics system for gameplay data // Computers&Graphics. – 2014. – №38. – P. 341-356.

[5] Бондарев А.Е. Оптимизационный анализ нестационарных пространственно-временных структур с применением методов визуализации / А.Е. Бондарев // Научная визуализация. – 2011. – том 3. – №2. – С. 1-11.

[6] Мильман И. Е. Анализ данных о деятельности кредитных организаций с использованием программы интерактивного визуального анализа многомерных данных / И. Е.Мильман [и др.] // Научная визуализация. – 2015. – том 7. – №1. – С. 45-64.

[7] Пилюгин В. В. Научная визуализация как метод анализа научных данных / В. В. Пилюгин [и др.] // Научная визуализация. – 2012. – том 4. – №4. – С. 56-70.

[8] Хачумов В. М. Разработка новых методов непрерывной идентификации и прогнозирования состояния динамических объектов на основе интеллектуального анализа данных / В. М. Хачумов, А. Н. Виноградов // Математические методы распознавания образов: 13-я Всероссиийская конференция. Ленинградская обл., г. Зеленогорск, 30 сентября - 6 октября 2007г. Сборник докладов. – 2007. – С. 548-550

Visual analytics of multidimensional dynamic data

Dmitry D. Popov, Igal E. Milman, Victor V. Pilyugin, Alexander A. Pasko

This work deals with a problem of analysis of a time variant object. Each object is characterized by a set of numerical parameters. The visualization method is used to conduct the analysis. Insights of interest for the analyst about the considered objects are obtained in several steps. At the first step, a geometric interpretation of the initial data is introduced. Then, the introduced mathematical model undergoes several transformations. These transformations correspond to solving the first problem of the visualization method, in particular obtaining visual representations of data. The next step is for the analyst to analyse the generated visual images and to interpret the results in terms of the considered objects.

We propose an algorithm for the problem solving. Developed interactive visualization software is described, which implements the proposed algorithm. We demonstrate how with this software the user can obtain insights regarding the creation and disappearance of object clusters and bunches and find invariants in the initial data changes.

Автоматизированная система сервисов обработки больших коллекций научных документов

© А.М. Елизаров © Е.К. Липачев © Ш.М. Хайдаров

Казанский (Приволжский) федеральный университет,
Казань

amelizarov@gmail.com elipachev@gmail.com 15jkeee@gmail.com

Аннотация

Представлена система сервисов автоматической обработки коллекций научных документов. Эти сервисы обеспечивают проверку соответствия документов принятым правилам формирования коллекций и их преобразование в установленные форматы; структурный анализ документов и извлечение метаданных, а также их интеграцию в научное информационное пространство. Система позволяет автоматически выполнять набор операций, который не реализуем за практически приемлемое время при традиционной «ручной» обработке электронного контента, и предназначена для больших коллекций научных документов.

1 Введение

Сегодня одной из актуальных проблем, стоящих перед человечеством, стала проблема накопления и последующей обработки огромных массивов данных. Под данными традиционно подразумевают различные необработанные информационные материалы, в том числе, данные различных наблюдений и научных экспериментов, персональные данные, а также различную статистическую информацию. По сведениям, приведенным в [1], уже в 2011 году каждый день создавалось около 15 PB новых данных, а за три года до этого момента времени человечество произвело информации больше, чем за всю историю своего существования до 2008 года, причем прирост данных происходил экспоненциально: это были и научные данные, и сведения о проведенных операциях-транзакциях, отчеты в социальных сетях и многое другое. Сегодня мировой объем данных увеличивается более чем в два раза каждые два года, а большие объемы данных (которые с 2008 года стали обозначать термином «большие данные» (Big Data)) открывают новые возможности и существенно влияют на развитие информационно-коммуникационных технологий (ИКТ).

Труды XVIII Международной конференции DAMDID/RCDL'2016 «Аналитика и управление данными в областях с интенсивным использованием данных», Ершово, 11-14 октября 2016

Традиционно термином Big Data обозначают наборы данных таких объёма и сложности, что стандартные инструменты работы с данными не способны осуществлять их обработку за время, приемлемое для практики [2]. Более широко этот термин можно трактовать как набор эффективных подходов, методов и инструментов обработки различных структурированных и неструктурированных данных большого объёма с целью получения приемлемых результатов в условиях непрерывного прироста данных [3]. Другими словами, термин «большие данные» характеризует совокупности данных, которые слишком велики по объему, характеризуются экспоненциальным ростом, не форматированы или не структурированы для анализа традиционными методами.

Не менее актуальна проблема учета значительного роста объёмов данных, получаемых, хранимых и обрабатываемых в ходе научной деятельности. В настоящее время благодаря внедрению ИКТ в научно-исследовательскую деятельность стало возможным при проведении новых исследований использовать весь корпус накопленных научных знаний. Это предполагает создание комплекса технологий, обеспечивающих оптимальное управление имеющимися знаниями, организацию эффективного доступа к ним, а также совместное и многократное использование новых видов структур знаний. В результате формируются разнообразные электронные научные коллекции и библиотеки, такие, например, как архивы научных журналов и отчетов, сборники научных трудов, диссертации и др. Они являются составной частью электронных научных библиотек и представляют собой наборы документов, имеющих различную структуру и разные форматы представления текстовых и графических материалов, библиографических списков, математической нотации. Эти различия затрудняют организацию информационных сервисов, опирающихся на машиноориентированную обработку информации (см., например, [4, 5]). Кроме того, в настоящее время значительно увеличивается объем данных, включаемых в коллекции, что в свою очередь создает дополнительные трудности при обработке научных Big Data. При управлении электронными научными коллекциями больших данных в полной мере

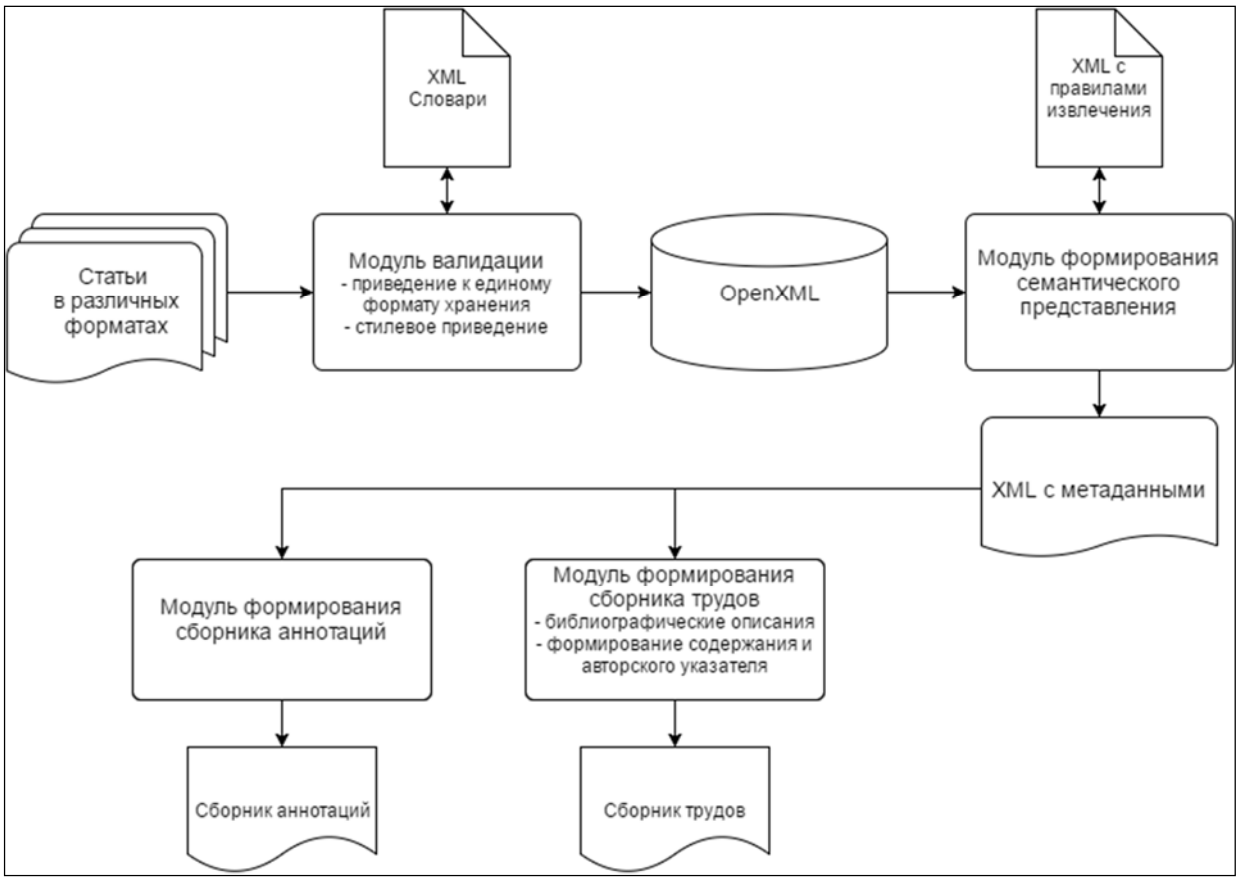


Рисунок 1 Архитектура системы

остаются актуальными, а также появляются новые задачи, в их числе: семантическая разметка, организация поиска, выделение метаданных, формирование тематических кластеров документов, сбор наукометрической информации, подготовка сборников материалов и др. (см., например, [6, 7]). Насущными становятся проблемы анализа и управления данными в различных областях с интенсивным использованием данных. Ниже представлена система сервисов автоматической обработки коллекций научных документов. С ее использованием проведена обработка материалов XI Всероссийского съезда по фундаментальным проблемам теоретической и прикладной механике (далее – Съезд), проведенного в Казани 20 – 24 августа 2015 г.: сформированы программа съезда, сборники аннотаций и трудов съезда (объемом более 1500 статей), а также соответствующая электронная коллекция.

2 Архитектура системы сервисов

На рис. 1 представлена архитектура созданной автоматизированной системы сервисов обработки больших коллекций научных документов. Она состоит из модулей, выполняющих следующие функции:

- извлечение метаданных из документов коллекции на основе анализа их структуры и форматов представления информации;
- автоматический выбор документов согласно установленному порядку, например, лексикографическому, по спискам авторов;
- извлечение блоков аннотаций из документов коллекции, подготовка алфавитного указателя и формирование сборника аннотаций;
- автоматическое формирование библиографического описания статьи коллекции с записью этой информации в блок колонтитулов документа;
- конвертация документов в pdf-формат в соответствии с установленными параметрами;
- формирование оригинал-макетов планируемых изданий с автоматической выборкой статей, расстановкой страниц, подготовкой алфавитного указателя и содержания;
- подготовка метаданных для экспорта в базы данных Российского индекса научного цитирования (РИНЦ).

3 Организация электронного хранилища

Машиноориентированная обработка электронных коллекций предполагает наличие семантической разметки их документов. Такая разметка частично присутствует в документах, использующих TEX-нотацию, при условии, что используются соответствующие макрокоманды (например, \title, \author, \abstract, \keywords) и стилевое окружение, характерное для каждой коллекции. В электронных научных коллекциях, представленных в офисных форматах (.doc, .docx и др.), а также .pdf, семантическая разметка отсутствует. Тем не менее, выполнить такую разметку можно в автоматическом режиме на основе информации о структурном строении каждого документа и особенностях его форматирования.

Прежде всего, коллекция разбивается на классы сходных по структуре документов, для каждого класса производится преобразование документов к семантическому представлению. С помощью набора паттернов регулярных выражений, специфичных для каждого класса документов, производится выделение информационных блоков (названия статьи, списка авторов, блока литературы и т. д.). В свою очередь, это дает возможность не только использовать семантические инструменты работы с электронным контентом, но и формировать в автоматическом режиме новые виды документов.

В хранилище организована навигация по названию, авторам и т. д. Реализация этих сервисов основана на структурном анализе документов в коллекции (см. раздел 5).

4 Сервис валидации и стилевого приведения

Под валидацией документов коллекции понимается процесс проверки наличия и расположения ключевых блоков (название статьи, список авторов, аффилиация, ключевые слова и т. д.), указанных в регламентируемых документах.

Сервис стилевого приведения реализует следующие шаги:

- единообразное представление названий статей; списка авторов (например, вместо Хайдаров Ш.М. записывается Ш.М. Хайдаров);
- единообразное представление аффилиации авторов, например, записи «КФУ», «К(П)ФУ», «Казанский университет», «Казанский (Приволжский) федеральный университет» и «Казанский федеральный университет» приводятся к единому виду «Казанский (Приволжский) федеральный университет»; для этого создается словарь синонимов;
- единообразное шрифтовое оформление разделов текста статей; происходит учет регистра при записи ключевых блоков, например, название статьи записывается прописными буквами;
- осуществляется выбор форматов рисунков, схем, диаграмм;
- производится набор математических формул и системы ссылок на них;
- списки литературы приводятся к выбранному формату библиографического описания.
- оформляются ссылки на поддержку исследований грантами, благодарности.

5 Формирование семантического представления коллекции на основе структурного анализа

Для извлечения метаданных статьи по характерным признакам (см. таблицу 1) определяются правила выделения блоков статьи. К таким признакам относятся стилевое оформление статей (шрифт, размер шрифта, выделение и т. д.). Кроме того, такие дополнительные признаки, как шаблонность текста (например, слово «Аннотация» перед блоком аннотаций или шаблонный вид электронной почты) и положение блока в тексте (например, документ начинается с названия статьи), позволяют повысить качество извлечения. В качестве таких признаков могут использоваться положение блока в документе, а также шрифт используемый в данном блоке (см., например, [8–11]). При структурном анализе коллекции научных документов Съезда использовался набор признаков, указанный в таблице 1.

Модуль реализован в виде PHP-скрипта, и его работа состоит из следующих шагов. Из файла статьи, хранящегося в формате docx, извлекается файл document.xml (см., например, [12]). Далее с использованием описания класса DOMDocument производится разбор этого файла. Для выделения блоков применяется метод getElementsByTagNameNS с параметром «w:p» (тег разметки абзаца в OpenXML). В результате получается список всех абзацев документа как объекта DOMNodeList. Полученный список последовательно проверяется на соответствие заданным правилам. В итоге для каждого документа (см. пример рис. 2) формируется его семантическое представление (рис. 3).

Для выделения семантических элементов разработан набор регулярных выражений, например, для выделения списка авторов используется выражение

/([А-ЯA-Z]\.(?:[А-ЯA-Z]\.)?\s[А-ЯA-Z][a-zа-я]+)(,\s)?(?1)?(,\s)?(?1)?/

Кроме того, проверяются наличие ключевых конструкций и их соответствие заданному формату. Результатом работы описываемого модуля является XML-документ, содержащий метаданные статей размечаемой коллекции.

Таблица 1 Характерные признаки для извлечения метаданных

Блок статьи	Признаки блока
Название статьи	Шрифт: Times New Roman, 12 пт, полужирный, выравнивание по центру. Положение: в начале документа
Список авторов	Шрифт: Times New Roman, 12 пт, выравнивание по центру Положение: после названия Шаблон имеют вид: И.О. Фамилия или И. Фамилия, перечисляются через запятую
Аффилиация	Шрифт: Times New Roman, 12 пт, курсив, выравнивание по центру. Положение: после списка авторов.
Электронная почта	Шрифт: Times New Roman, 9 пт, выравнивание по центру Положение: после аффилиации Шаблон содержат символ @ и имеют заданный вид
Аннотация	Шрифт: Times New Roman, 9 пт, выравнивание по ширине Положение: после адреса электронной почты Шаблон начинается со слова «Аннотация».

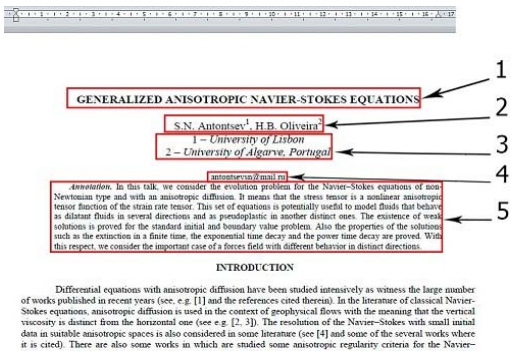


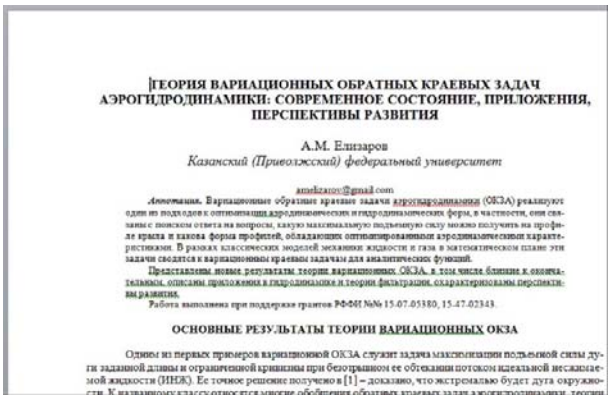
Рисунок 2 Пример статьи, где 1 – название, 2 – блок списка авторов, 3 – блок аффилиации, 4 – блок электронной почты и 5 – блок аннотации. Стилевое оформлении соответствует таблице 1

```
<?xml version="1.0" encoding="UTF-8"?>
<articles>
 <article>
 <eachauthors>
  <author>S.N. Antontsev</author>
  <author>H.B. Oliveira</author>
  <workplace>University of Lisbon; University of
Algarve, Portugal</workplace>
  <mails>antontsevsn@mail.ru</mails>
 </eachauthors>
 <artTitles>
  <artTitle>GENERALIZED ANISOTROPIC
NAVIER-STOKES EQUATIONS</artTitle>
 </artTitles>
 <abstracts>
  <abstract>&amp;lt;p style=&amp;quot;text-indent:
        20px;&amp;quot;&amp;gt;In this talk, we consider
        the evolution problem for the Navier–
        Stokes equations of non-Newtonian type
        and with an anisotropic diffusion.
...
&amp;lt;/p&amp;gt;</abstract>
 </abstracts>
 <files>
  <furl>F:\Desktop\doc/00001.pdf</furl>
 </files>
</article>
…
</articles>
```

Рисунок 3 Фрагмент сгенерированного XML-файла

6 Модуль формирования оригинал-макета научного издания

Этот модуль позволяет в автоматическом режиме подготовить из файлов электронной коллекции оригинал-макет научного издания (сборник материалов, труды и т. д.). Порядок размещения статей определяется семантическим представлением коллекции, хранящемся в XML-файле (см. раздел 5). Алгоритм реализован в виде макроса VBA и включает следующие шаги: сначала для задания диапазона страниц статей определяются счетчики начальной и конечной страниц и задаются их начальные значения (см. рис. 4–6). Далее последовательно открываются документы коллекции в соответствии с порядком, заданным в XML-файле в соответствии с правилами извлечения. Вычисляются начальные и конечные страницы, после чего формируется библиографическое описание статьи, которое записывается в колонтитул данного документа. Полученный документ конвертируется в PDF-формат. Также библиографическое описание сохраняется в XML-файле. На рис. 6 приведен фрагмент кода, выполняющий описанные операции.


ТЕОРИЯ ВАРИАЦИОННЫХ ОБРАТНЫХ КРАЕВЫХ ЗАДАЧ АЭРОГИДРОДИНАМИКИ: СОВРЕМЕННОЕ СОСТОЯНИЕ, ПРИЛОЖЕНИЯ, ПЕРСПЕКТИВЫ РАЗВИТИЯ

А.М. Елизаров

Казанский (Приволжский) федеральный университет

amelizarov@gmail.com

Аннотация. Вариационные обратные краевые задачи аэрогидродинамики (ОКЗА) реализуют один из подходов к оптимизации аэродинамических и гидродинамических форм, в частности, они связаны с поиском ответа на вопросы, какую максимальную подъемную силу можно получить на профиле крыла и какова форма профилей, обладающих оптимизированными аэродинамическими характеристиками. В рамках классических моделей механики жидкости и газа в математическом плане эти задачи сводятся к вариационным краевым задачам для аналитических функций.

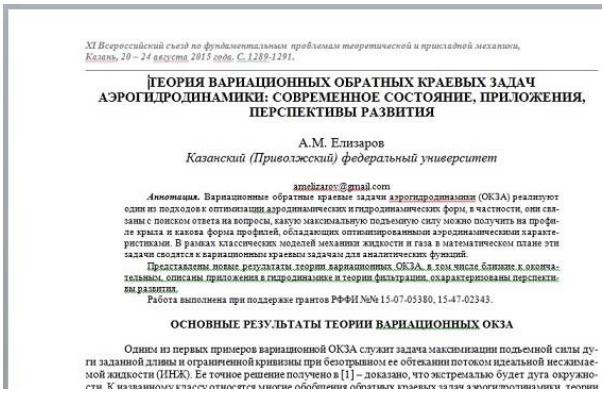
Представлены новые результаты теории вариационных ОКЗА, в том числе близкие к окончательным, описаны приложения в гидродинамике и теории фильтрации, охарактеризованы перспективы развития.

Работа выполнена при поддержке грантов РФФИ №№ 15-07-05380, 15-47-02343.

ОСНОВНЫЕ РЕЗУЛЬТАТЫ ТЕОРИИ ВАРИАЦИОННЫХ ОКЗА

Одним из первых примеров вариационной ОКЗА служит задача максимизации подъемной силы дуги заданной длины и ограниченной кривизны при безотрывном ее обтекании потоком идеальной несжимаемой жидкости (ИНЖ). Ее точное решение получено в [1] – доказано, что экстремалью будет дуга окружности. К названному классу относятся многие обобщения обратных краевых задач аэрогидродинамики, теории

Рисунок 4 Фрагмент документа до обработки модулем



XI Всероссийский съезд по фундаментальным проблемам теоретической и прикладной механики, Казань, 20 – 24 августа 2015 года. С. 1289-1291.

ТЕОРИЯ ВАРИАЦИОННЫХ ОБРАТНЫХ КРАЕВЫХ ЗАДАЧ АЭРОГИДРОДИНАМИКИ: СОВРЕМЕННОЕ СОСТОЯНИЕ, ПРИЛОЖЕНИЯ, ПЕРСПЕКТИВЫ РАЗВИТИЯ

А.М. Елизаров

Казанский (Приволжский) федеральный университет

amelizarov@gmail.com

Аннотация. Вариационные обратные краевые задачи аэрогидродинамики (ОКЗА) реализуют один из подходов к оптимизации аэродинамических и гидродинамических форм, в частности, они связаны с поиском ответа на вопросы, какую максимальную подъемную силу можно получить на профиле крыла и какова форма профилей, обладающих оптимизированными аэродинамическими характеристиками. В рамках классических моделей механики жидкости и газа в математическом плане эти задачи сводятся к вариационным краевым задачам для аналитических функций.

Представлены новые результаты теории вариационных ОКЗА, в том числе близкие к окончательным, описаны приложения в гидродинамике и теории фильтрации, охарактеризованы перспективы развития.

Работа выполнена при поддержке грантов РФФИ №№ 15-07-05380, 15-47-02343.

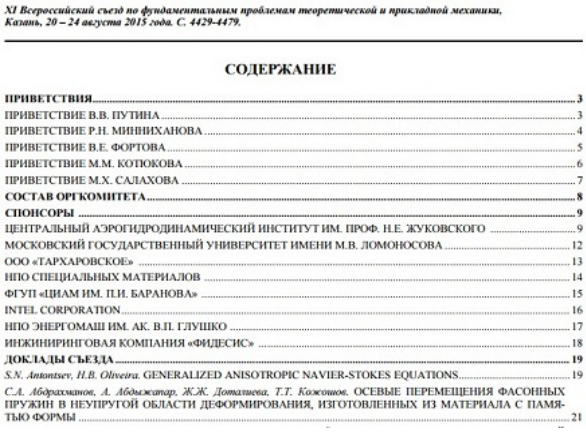
ОСНОВНЫЕ РЕЗУЛЬТАТЫ ТЕОРИИ ВАРИАЦИОННЫХ ОКЗА

Одним из первых примеров вариационной ОКЗА служит задача максимизации подъемной силы дуги заданной длины и ограниченной кривизны при безотрывном ее обтекании потоком идеальной несжимаемой жидкости (ИНЖ). Ее точное решение получено в [1] – доказано, что экстремалью будет дуга окружно-

Рисунок 5 Фрагмент документа после обработки модулем (создан колонтитул с выходными данными)

```
Sub Макрос1()
'
' Макрос1 Макрос
'
'
 Application.ScreenUpdating = False
 StartPage = 4367
 sPath = "F:\Desktop\doc1\"
 sFile = Dir(sPath & "*.docx")
 While sFile <> ""
  With Documents.Open(sPath & sFile)
  ActiveDocument.PageSetup.HeaderDistance = CentimetersToPoints(1)
  deleteAllHeaders_Footers
  ActiveDocu-ment.PageSetup.DifferentFirstPageHeaderFooter = True
  ActiveWindow.ActivePane.View.SeekView = wdSeekFirstPageHeader
  ActiveDocument.Repaginate
  EndPage = StartPage + ActiveDocu-ment.BuiltInDocumentProperties(wdPropertyPages) - 1
  Selection.Font.Name = "Times New Roman"
  Selection.Font.Size = 9
  Selection.Font.Bold = wdToggle
  Selection.Font.Italic = wdToggle
  Selection.TypeText Text:="XI Всероссийский съезд по фундаментальным проблемам теоретической и прикладной механики,"
  Selection.TypeParagraph
  Selection.TypeText Text:="Казань, 20 – 24 авгу-ста 2015 года. С. "
  Selection.TypeText StartPage
  Selection.TypeText Text:="-"
  Selection.TypeText EndPage
  Selection.TypeText Text:="."
  Selection.TypeParagraph
  Selection.InlineShapes.AddHorizontalLineStandard
  Selection.MoveLeft Unit:=wdCharacter, Count:=2, Extend:=wdExtend
  Selection.InlineShapes(1).Fill.Visible = msoTrue
  Selection.InlineShapes(1).Fill.Solid
  Selection.InlineShapes(1).Fill.ForeColor.RGB = RGB(0, 0, 0)
  Selection.InlineShapes(1).Fill.Transparency = 0#
  Selec-tion.InlineShapes(1).HorizontalLineFormat.WidthType = _
   wdHorizontalLinePercentWidth
  Selec-tion.InlineShapes(1).HorizontalLineFormat.PercentWidth = 100
  Selection.InlineShapes(1).Height = 1
  Selec-tion.InlineShapes(1).HorizontalLineFormat.NoShade = True
  Selection.InlineShapes(1).HorizontalLineFormat.Alignment = _
   wdHorizontalLineAlignCenter
  ActiveWindow.ActivePane.View.SeekView = wdSeekMainDocument
  ActiveWindow.ActivePane.View.SeekView = wdSeekFirstPageFooter
  Selection.Fields.Add Range:=Selection.Range, Type:=wdFieldEmpty, PreserveFormatting:=False
  Selection.TypeText Text:="PAGE"
  Selection.Fields.Update
  Selection.Fields.ToggleShowCodes
  Selection.Font.Name = "Times New Roman"
  Selection.Font.Size = 12
  Selection.ParagraphFormat.Alignment = wdAlign-ParagraphRight
  ActiveWindow.ActivePane.View.SeekView = wdSeekMainDocument
  With ActiveDocu-ment.Sections(1).Footers(wdHeaderFooterPrimary).PageNumbers
   .RestartNumberingAtSection = True
   .StartingNumber = StartPage
   .Add wdAlignPageNumberRight, False
  End With
  StartPage = EndPage + 1
  ActiveDocument.ExportAsFixedFormat Output-FileName:=sPath & Replace(sFile, "docx", "pdf"), Export-Format _
  :=wdExportFormatPDF, OpenAfterExport:=False, OptimizeFor:=_
  wdExportOptimizeForPrint, Range:=wdExportAllDocument, From:=1, To:=1, _
  Item:=wdExportDocumentContent, Includ-eDocProps:=True, KeepIRM:=True, _
  CreateBookmarks:=wdExportCreateNoBookmarks, DocStructureTags:=True, _
  BitmapMissingFonts:=True, UseI-SO19005_1:=False
  On Error Resume Next
  ActiveDocument.Close (True)
  End With
  sFile = Dir
 Wend
End Sub
```

Рисунок 6 Фрагмент кода формирования библиографического описания

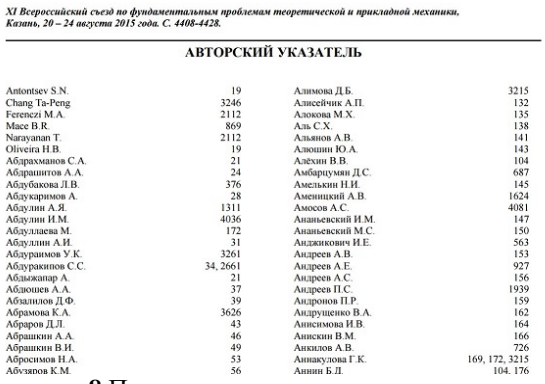


XI Всероссийский съезд по фундаментальным проблемам теоретической и прикладной механики, Казань, 20 – 24 августа 2015 года. С. 4429-4479.

СОДЕРЖАНИЕ

Рисунок 7 Пример автоматически сгенерированного содержания сборника трудов Съезда

После обработки всех документов коллекции формируются содержание издания и авторский указатель. При этом используются данные, сохраненные в XML-файле на этапе формирования колонтитулов. На рисунках 7 и 8 приведены автоматически сформированные содержание издания и авторский указатель.



XI Всероссийский съезд по фундаментальным проблемам теоретической и прикладной механики, Казань, 20 – 24 августа 2015 года. С. 4408-4428.

АВТОРСКИЙ УКАЗАТЕЛЬ

Antontsev S.N.	19	Алимова Д.Б.	3215
Chang Ta-Peng	3246	Алисейчик А.П.	132
Ferenczi M.A.	2112	Алокова М.Х.	135
Mace B.R.	869	Аль С.Х.	138
Narayanan T.	2112	Альянов А.В.	141
Oliveira H.B.	19	Алюшин Ю.А.	143
Абдрахманов С.А.	21	Алёхин В.В.	104
Абдрашитов А.А.	24	Амбарцумян Д.С.	687
Абдубакова Л.В.	376	Амелькин Н.И.	145
Абдукаримов А.	28	Аменицкий А.В.	1624
Абдулин А.Я.	1311	Амосов А.С.	4081
Абдулин И.М.	4036	Ананьевский И.М.	147
Абдуллаева М.	172	Ананьевский М.С.	150
Абдуллин А.И.	31	Анджикович И.Е.	563
Абдураимов У.К.	3261	Андреев А.В.	153
Абдуракипов С.С.	34, 2661	Андреев А.Е.	927
Абдыжапар А.	21	Андреев А.С.	156
Абдюшев А.А.	37	Андреев П.С.	1939
Абилилов Д.Ф.	39	Андронов П.Р.	159
Абрамова К.А.	3626	Андрущенко В.А.	162
Абраров Д.Л.	43	Анисимова И.В.	164
Абрашкин А.А.	46	Анискин В.М.	166
Абрашкин В.И.	49	Анкилов А.В.	726
Абросимов Н.А.	53	Аннакулова Г.К.	169, 172, 3215
Абузяров К.М.	56	Аннин Б.Д.	104. 176

Рисунок 8 Пример автоматически сгенерированного авторского указателя сборника трудов Съезда

7 Сервис извлечения библиографических метаданных и загрузки в РИНЦ

Алгоритм извлечения библиографических метаданных и загрузки их в РИНЦ состоит из следующих шагов (проиллюстрированных на примере материалов Съезда):

1. из оригинал-макета сборника трудов извлечены библиографические описания каждой публикации;
2. соответствующий скрипт находит в документе блок библиографических описаний и с помощью регулярных выражений разделяет их по видам изданий (например, отличительным признаком библиографического описания статьи является наличие знака //);
3. проводится разбор основных метаданных – выделяются список авторов, названия статей, изданий и т. д.;
4. с помощью разработанного веб-приложения генерируется XML-файл в соответствии с правилами РИНЦ, содержащий набор метаданных публикации.

Заключение

Предложен метод автоматической обработки больших коллекций физико-математических документов, включающий их валидацию и семантический анализ, извлечение метаданных, подготовку различных видов оригинал-макетов научных изданий. Метод позволяет выполнять автоматическую обработку больших коллекций электронных документов с набором операций, который не реализуем при традиционной «ручной» работе с электронным контентом.

Приведен пример успешной его реализации при организации XI Всероссийского съезда по фундаментальным проблемам теоретической и прикладной механики (Казань, 20–24 августа 2015 г.).

Благодарности

Работа выполнена при финансовой поддержке РФФИ (проекты №№ 15-07-08522, 15-47-02472).

Литература

[1] IBM’s Top Storage Predictions for 2011, January 2011, StorageNewsletter.com.

[2] MIKE2.0. The open source standard for information management. Big Data definition. http://mike2.openmethodology.org/wiki/Big_Data_Definition

[3] A Manyika J., M. Chui, B. Brown, J. Bughin, R. Dobbs, C. Roxburgh, A. H. Byers. Big data: The next frontier for innovation, competition, and productivity: McKinsey Global Institute Report, 2011. http://www.mckinsey.com/insights/mgi/research/technology_and_innovation/big_data_the_next_frontier_for_innovation

[4] P. J. Olver. Journals in flux. Notices Amer. Math. Soc., V. 58 (8), 2011, p. 1124-1126.

[5] С. А. Афонин, А. В. Бахтин, В. Ю. Бухонов, В. А. Васенин, Г. М. Ганкин, А. Э. Гаспарянц, Д. Д. Голомазов, А. А. Иткес, А. С. Козицын, И. Н. Тумайкин, К. А. Шапченко. Интеллектуальная система тематического исследования научно-технической информации (ИСТИНА). М.: Изд-во Московского ун-та, 2014, 262 с.

[6] А. М. Елизаров, Н. Г. Жильцов, А. В. Кириллович, Е. К. Липачёв. Семантическое аннотирование в системе управления физико-математическим контентом. Науч. сервис в сети Интернет: труды XVII Всерос. науч. конф. (21–26 сентября 2015 г., г. Новороссийск), М.: ИПМ им. М.В. Келдыша, с. 98-103, 2015.

[7] А. М. Елизаров, Н. Г. Жильцов, А. В. Кириллович, Е. К. Липачёв. Терминологическое аннотирование и рекомендательный сервис в системе управления физико-математическим контентом. Труды XVII Межд. конф. DAMDID / RCDL’2015 «Аналитика и управление данными в областях с интенсивным использованием данных». Обнинск: ИАТЭ НИЯУ МИФИ, с. 347-350, 2015.

[8] Xiaonan Lu, Brewster Kahle, James Z. Wang and C. Lee Giles. A metadata generation system for scanned scientific volumes. Joint Conference on Digital Libraries, June 16–20, 2008, Pittsburgh, Pennsylvania, p. 167-176, 2008.

[9] J. Chen, H. Chen. A structured information extraction algorithm for scientific papers based on feature rules learning. Journal of Software, Vol. 8(1), p. 55-62, 2013. http://www.jsoftware.us/vol8/jsw0801-08.pdf

[10] D. Tkaczyk, B. Tarnawski, L. Bolikowski. Structured affiliations extraction from scientific literature. D-Lib Magazine, V. 21 (11/12), 2015. http://www.dlib.org/dlib/november15/tkaczyk/11tkaczyk.html

[11] А.М. Елизаров, Е.К. Липачёв, Ш.М. Хайдаров. Автоматизированная система структурной и семантической обработки физико-математического контента. Ученые записки Института социально-гуманитарных знаний, № 1 (14), с. 210-215, 2016.

[12] Standard ECMA-376: Office Open XML File Formats. http://www.ecma-international.org/publications/standards/Ecma-376.htm

Automated system of services for processing of large collections of scientific documents

Alexander M. Elizarov, Evgeny K. Lipachev, Shamil M. Khaydarov

This paper presents a system of services for the automated processing of collections of scientific documents. These services provide verification of document compliance to the accepted rules of formation of collections and their conversion to the established formats; structural analysis of documents and extraction of metadata, as well as their integration into the scientific information space. The system allows to automatically perform a set of operations that cannot be realized for acceptable time with the traditional manual processing of electronic content. It is designed for the large collections of scientific documents.

Системы управления обучением

Learning Management Systems

Применение нечётких когнитивных карт для моделирования поведения пользователей системы дистанционного обучения

В.С. Киреев

Национальный исследовательский ядерный университет «МИФИ»,
Москва, Российская Федерация
vskireev@mephi.ru

Аннотация

Оптимизация систем дистанционного обучения (LMS) является одной из актуальных задач в условиях роста объёмов представляемого в них контента и увеличения количества слушателей дистанционных курсов. Оптимизация в большинстве случаев основывается на анализе данных логов LMS и выявлении паттернов поведения пользователей по отношению к контенту. Данная статья посвящена описанию подхода к моделированию пользовательского поведения в системе дистанционного обучения на основе подхода, включающего нечёткие когнитивные карты (FCM). Предлагаемая модель описывает взаимодействие пользователей с контентом в системе и может быть использована для прогнозирования реакции пользователей на обучающие, контрольные и практические элементы. Полученная когнитивная карта протестирована и уточнена с помощью данных системы ИНФОМИФИСТ, использующейся на ряде факультетов НИЯУ МИФИ для поддержки учебного процесса уже более 9 лет.

1 Введение

В настоящее время дистанционное образование (e-learning) вызывает повышенный интерес как в корпоративном секторе, с массовым внедрением концепции корпоративного университета, так и среди классических образовательных учреждений [4]. Также следует отметить платформы, посвящённые обучению как таковому и поддерживающие концепцию МООК, например, Coursera, EdX и т.д. Данная парадигма позволяет повысить эффективность процесса обучения, снизить организационные и производственные издержки, автоматизировать процесс передачи знаний и получить дополнительный источник информации о качестве получаемого слушателями образования и их поведении. Системы дистанционного образования представлены как корпоративными решениями, так и решениями с открытым кодом. Среди последних, наибольшей популярностью пользуется система Moodle, которая применяется во многих учебных заведениях в качестве фреймворка для собственных программных решений. Дистанционное образование в последние годы характеризуется накоплением большого объёма данных и востребованностью для его обработки методов интеллектуального анализа (E-learning data mining), поэтому проблема повышения эффективности обучения слушателей, за счёт оптимизации обучающего контента, является крайне актуальной.

Труды XVIII Международной конференции DAMDID/RCDL'2016 «Аналитика и управление данными в областях с интенсивным использованием данных», Ершово, 11-14 октября 2016

2 Современные подходы к решению проблемы

Большинство подходов к оптимизации контента интернет-ресурсов вообще и систем дистанционного образования в частности, заключается в выявлении паттернов поведения пользователей на основе анализа их действий в системе (Web Mining) [2,3]. Данный анализ, в основном, производится с использованием методов кластеризации и классификации. Кроме того, используются методы поиска ассоциативных правил, секвенциальный анализ и текстовый (контент) анализ (см. Рис. 1).

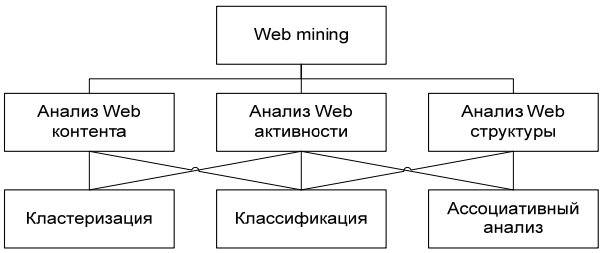


Рисунок 1 Актуальные направления интеллектуального анализа интернет-ресурсов

Кроме этого в последние годы развивается направление когнитивной визуализации [10], позволяющей описывать учебную траекторию слушателя по имеющимся логам LMS. Перечисленные методы позволяют выявить локальные особенности поведения пользователей, однако, возможность обобщения полученных результатов на процесс взаимодействия с LMS в

целом, возникает достаточно редко [5]. Таким образом, для выявления слабых мест представленного в системе контента возникает необходимость проведения повторных исследований, в большинстве случаев достаточно трудоёмких. Создатель курса в силу различных причин не всегда имеет возможность реализовать данный вариант, и присутствующие в системе данные оказываются невостребованными. Таким образом, выявление требований к контенту на фундаментальной основе является необходимым и обоснованным. Эти требования должны закладываться на первом этапе, при разработке курса для создания более совершенного контента. С этой целью предлагается разработать обобщённую модель, которая позволит определить оптимальные количественно-качественные формы обучающего контента.

3 Нечёткие когнитивные карты

Когнитивные карты являются одним из инструментов представления слабо формализуемых предметных областей, в особенности в экономике, политической и военной сферах. Данный подход был предложен Аксельродом в работе 1976 года, посвящённой моделированию политической сферы [7]. Когнитивная карта представляет собой знаковый ориентированный граф, в котором вершинами представляются сущности, концепции, факторы, цели и события, а дугами задаётся их влияние друг на друга. Влияние или воздействие характеризуется некоторой пороговой функцией, которая может определяться различными способами. Функция получается на основании экспертной оценки, которая первоначально задаётся в лингвистической форме. Впоследствии Кошко [1] предложил расширение данной парадигмы за счёт введения нечёткости, что в большей степени отражает разброс мнений экспертов при оценке воздействия одних факторов на другие. В качестве нечётких чисел чаще всего используются треугольные числа.

В целом, задача определения состояния вершин (концептов) когнитивной карты сводится к расчётам в соответствии с формулой (см. формулу 1):

$$A_i(k+1) = f\left(A_i(k) + \sum_{j \neq i,\, j=1}^{N} A_j(k) W_{ji} \right), \quad (1)$$

где $A_i(k+1)$ – новое состояние вершины, $A_i(k)$ – предыдущее состояние, W_{ji} – матрица весов, f – пороговая функция.

Процесс расчёта является итеративным - после задания начальных состояний вершин значения состояний пересчитываются до тех пор, пока разница между текущими и предшествующими состояниями не окажется меньше некоторого заданного значения ε.

На сегодняшний день в процессе управления сложными системами часто используются когнитивные карты [6] разной степени формализации на различных этапах поиска решений в слабоструктурированных проблемных областях, особенно в социальной и экономической сфере. На основе когнитивных карт разрабатываются методики генерации и верификации карт, поддерживающие этап формирования общего представления знаний о ситуации. На данном этапе разработка когнитивных карт направлена на визуальное представление проблемы [11,12] для объяснения действий субъекта, опираясь на анализ его точки зрения. В этом случае адекватность карты подтверждается самим субъектом. Одним из актуальных направлений развития когнитивных карт являются фреймовые когнитивные карты [9].

4 Предлагаемый подход

В качестве модели взаимодействия пользователя с системой LMS предлагается нечёткая когнитивная карта, описывающая воздействие на эффективность освоения (ИК) слушателем курса набора концептов, характеризующих контент как с дидактической, так и системной точек зрения. В частности, изучаемый курс представляет собой совокупность модулей (МК), результативность освоения которых влияет на концепт ИК. Отдельный модуль представляется совокупностью статического и интерактивного контента. Статический обучающий контент (ОК) включает в себя конспекты и презентации лекций, вспомогательные методические материалы и т.д. Интерактивный контент включает в себя чистый контролирующий компонент, такой как тесты (ТК), и обучающий практико-ориентированный компонент (ПК), такой как лабораторные работы и тренажёры, выполняемые в системе (например, SCORM-пакеты).

Среди других сущностей можно отметить: Взаимодействие с системой (ВС), Количество входов (КВ), Проведённое в системе время (ПВ), Качество обратной связи (КО), Количество новых тем, созданных пользователем (КН), Количество сообщений, оставленных пользователем (КС), Успешность освоения текущего курса (К1), Успешность освоения других курсов (KN), Результаты других слушателей (РС), Итоговая оценка за курс (ОИ), Траектория обращения пользователя к модулям, и контенту внутри них, соответствующая естественной последовательности освоения курса (ПО), Успешность освоения модуля курса (МК), Рекомендуемые источники (РИ), Обучающий контент (ОК), Практико-ориентированный компонент (ПК), Контрольный компонент (ТК), Количество материалов, Количество тестов (КТ), Количество попыток прохождения теста (КП), Оценка за тест (ОТ), Время проведённое в тесте (ВТ) (см. Рис. 2).

Путём опроса 5 экспертов – разработчиков дистанционных курсов были определены веса дуг, для этого использовались однотипные лингвистические шкалы, рассчитанные по 5-балльной шкале, где 1 балл означает низкий уровень

влияния, а 5- максимально высокий. Далее оценки экспертов были согласованы с помощью метода Саати и осуществлена их фаззификация. В качестве пороговой функции для состояний концептов использовалась функция единичного скачка (см. формулу 2):

$$f(x)=\begin{cases}1, x \geq 0\\ 0, x \leq 0\end{cases}, \qquad (2)$$

Предложенная модель была протестирована с помощью свободно распространяемого приложения Fcmapper на нескольких сценариях, зависящих от сложности практико-ориентированного, самостоятельного и контрольного компонентов, и количества соответствующих элементов, для поиска равновесного состояния (см. Рис. 3).

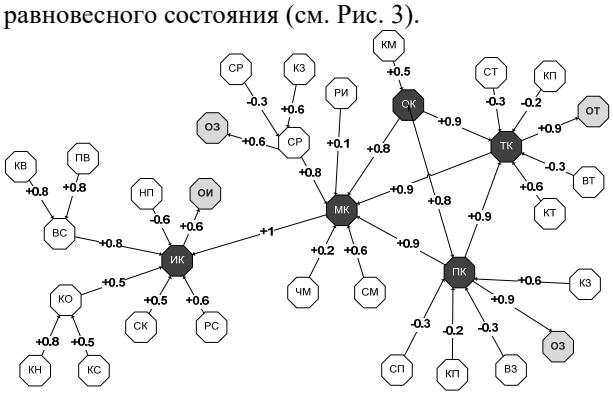


Рисунок 2 Когнитивная карта взаимодействия пользователей с системой дистанционного образования

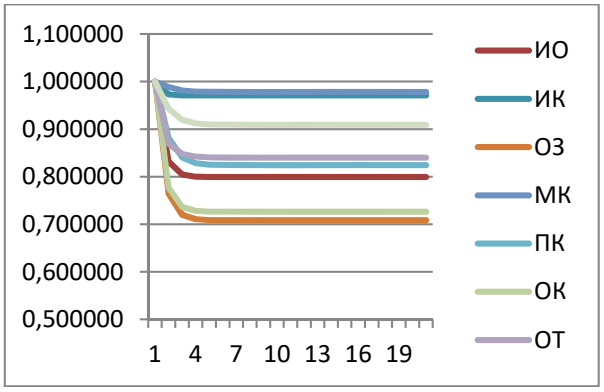


Рисунок 3 Пример результатов использования сценария с «лёгким» уровнем сложности компонентов курса

На графике представлены основные концепты, включая эффективность освоения курса (ИК), и можно видеть, что в условиях «низкой» сложности значение концепта ИК достигает вполне приемлемой величины – 0,8. Индикаторные показатели – оценки за компоненты курса также укладываются в нормативные значения, в том числе по системе ECTS.

5 Тестирование предлагаемой модели

Верификация когнитивных карт представляет собой нетривиальную задачу и чаще всего решается двумя различными способами [8]: 1) проверкой полученных значений в вершинах или модели в целом с помощью альтернативных подходов и методов – математико-статистических, таких как метод Монте-Карло, или приёмами имитационного моделирования и т.д.; 2) проверкой вывода каждой дуги на реальных исторических данных. Построенная модель была протестирована на реальных данных, полученных из системы ИНФОМИФИСТ (см. Рис. 4), используемой для поддержки учебного процесса в НИЯУ МИФИ с 2007 года на факультете Экономики и Управления, факультете Кибернетики и Информационной безопасности, а с 2015 года в Бизнес школе. За это время в системе было размещено более 100 различных курсов, обучение которым прошли около 15 тысяч слушателей. ИНФОМИФИСТ разработан на основе системы дистанционного обучения с открытым кодом Moodle, которая позволяет управлять контентом, пользователями и осуществлять мониторинг их действий. Таким образом, был накоплен большой объём информации о поведении пользователей, который был проанализирован с целью уточнения параметров разработанной когнитивной карты. Журнал событий системы содержит более 3 миллионов записей.

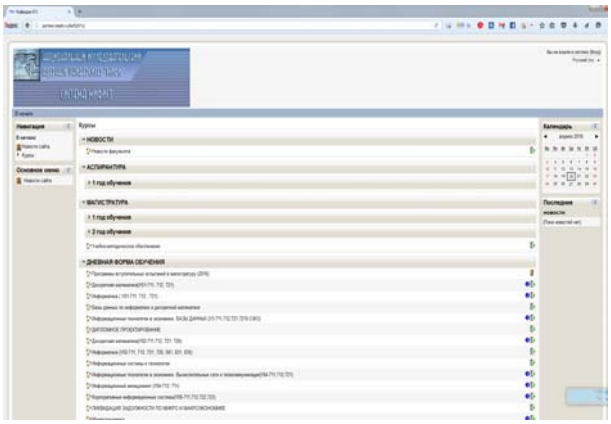


Рисунок 4 Главная страница системы ИНФОМИФИСТ

Для проведения анализа были использованы данные системы, соответствующие учебному плану магистерского направления подготовки 38.04.01 «Экономика». С помощью корреляционно-регрессионного подхода были проведены оценки зависимости ряда предложенных в модели сущностей, таких как индикаторные элементы (оценки за практический компонент и другие оценки) от сущностей второго порядка – количество лабораторных работ в SCORM (КЗ), количество сделанных попыток (КП), общее затраченное время (ВЗ) и т.д. (см. формулу 3):

$$Y^{i}_{OT} = F(X^{i}_{СП}, X^{i}_{КП}, X^{i}_{КЗ}, X^{i}_{ВЗ}) + \varepsilon_i \qquad (3)$$

В качестве модельной функции регрессии была использована квадратичная функция. В качестве исходных данных по каждому компоненту была взята агрегация показателей по каждому слушателю.

Например, по практическому компоненту: сумма оценок за все SCORM работы, сумма минут, потраченных на все задания, сумма количества попыток на каждое задание, количество всех заданий.

Рисунок 5 Результаты корреляционно-регрессионного анализа на пример курса «Маркетинг инноваций»

Полученные результаты (см. на Рис. 5) свидетельствуют о соответствии в целом оцененных экспертами весов дуг модели и построенной регрессии. Однако ряд значений весов требуется пересмотреть, например, воздействие проведённого времени в тесте. Кроме этого, для повышения точности работы модели планируется выделить дополнительные сущности и факторы, представляющие компоненты курса более точно, в том числе по обучающему контенту.

6 Заключение

Когнитивные карты позволяют моделировать слабо формализованные предметные области для повышения качества прогнозов, создания возможных сценариев развития ситуации. В данной статье обсуждается возможность использования нечётких когнитивных карт в качестве основы модели поведения пользователей в процессе дистанционного обучения с помощью LMS. Планируются дальнейшие исследования с целью уточнения параметров построенной когнитивной карты в части нечётких функций, описывающих взаимное влияние сущностей карты, а также весовых значений дуг.

Литература

[1] B. Kosko, Fuzzy Cognitive Maps, Int. J. of Man-Machine Studies, 24, 1986, p. 65-75.

[2] C. Romero, S. Ventura, & E. Garcia. (2008). Data mining in course management systems: Moodle case study and tutorial. Computers & Education, 51(1), 368-384.

[3] E. Garcia, C. Romero, S. Ventura, S., & C. de Castro. (2011). A collaborative educational association rule mining tool. Internet and Higher Education, 14(2), 77-88.

[4] E-Learning Market Trends & Forecast 2014 – 2016 URL: https://www.docebo.com/landing/contactform/elearning-market-trends-and-forecast-2014-2016-docebo-report.pdf (дата обращения: 15.05.2016).

[5] Kavita D. Satokar, Prof..S.z.Gawali, “Web Search Result Personalization using Web Mining”, International Journal of Computer Applications (0975 - 8887), Volume 2 - No.5, June 2010.

[6] O.D. Ntarlas, P. P Groumpos. A survey on Applications of fuzzy cognitive maps in business and management // Вестник УГАТУ = Vestnik UGATU. 2014. №5. URL: http://cyberleninka.ru/article/n/a-survey-on-applications-of-fuzzy-cognitive-maps-in-business-and-management (дата обращения: 15.05.2016).

[7] R. Axelrod, Structure of Decision, The Cognitive Maps of Political Elite, Princeton University Press, 1976.

[8] А. А. Кулинич, Компьютерные системы моделирования когнитивных карт: подходы и методы, Проблемы управления, 2010, № 3, 2–16.

[9] А.А. Кулинич Семиотические когнитивные карты (фреймовая модель) / Труды XII Всероссийского совещания по проблема управления (ВСПУ-2014, Москва). М.: Ипу РАН, 2014. С. 4152-4164.

[10] В.А. Углев, Т.М. Ковалева Когнитивная визуализация как инструмент сопровождения индивидуального обучения // Наука и образование: научное издание МГТУ им. Н.Э. Баумана. 2014. №3. URL: http://cyberleninka.ru/article/n/kognitivnaya-vizualizatsiya-kak-instrument-soprovozhdeniya-individualnogo-obucheniya (дата обращения: 15.05.2016).

[11] В.П. Карелин Модели и методы представления знаний и выработки решений в интеллектуальных информационных системах с нечёткой логикой // Вестник ТИУиЭ. 2014. №1 (19). URL: http://cyberleninka.ru/article/n/modeli-i-metody-predstavleniya-znaniy-i-vyrabotki-resheniy-v-intellektualnyh-informatsionnyh-sistemah-c-nechyotkoy-logikoy (дата обращения: 15.05.2016).

[12] Л.А. Гинис Развитие инструментария когнитивного моделирования для исследования сложных систем // ИВД. 2013. №3 (26). URL: http://cyberleninka.ru/article/n/razvitie-instrumentariya-kognitivnogo-modelirovaniya-dlya-issledovaniya-slozhnyh-sistem (дата обращения: 18.05.2016).

Application of fuzzy cognitive maps in simulation of the LMS users' behavior

Vasiliy S. Kireev

Optimization of learning management systems (LMS) is one of the urgent tasks in the face of rising volumes of content represented in them t and increasing the number of listeners of online courses. Optimization in most cases is based on the analysis of LMS log data, to identify patterns of user behavior relative to content. This article describes the approach to modeling user behavior for systems of distance learning-based approach involving fuzzy cognitive maps (FCM). The proposed model describes the user interaction with the content in the system and can be used to predict how users react to the training, control and practical elements. The resulting cognitive map is tested and refined using data system INFOMEPHIST applied in a number of faculties of the MEPhI to support the educational process for more than 9 years.

Мониторинг потребностей рынка труда в выпускниках вузов на основе аналитики с интенсивным использованием данных*

© П.В. Зрелов[1, 2] © В.В. Кореньков[1, 2, 3] © Н. А. Кутовский[1, 2]
© А.Ш. Петросян[1, 2] © Б. Д. Румянцев[1] © Р. Н. Семенов[1, 2]
© И. А. Филозова[1, 3]

[1]Лаборатория информационных технологий, ОИЯИ,
[3]Государственный университет «Дубна»,
Дубна
[2]Российский экономический университет им. Г.В. Плеханова,
Москва

zrelov@jinr.ru korenkov@jinr.ru kut@jinr.ru
virthead@jinr.ru bdrum@jinr.ru roman@jinr.ru
fia@jinr.ru

Аннотация

В работе рассматривается проблема объективной оценки состояния рынка труда и подготовки выпускников, соответствующих ожиданиям работодателей, дается описание автоматизированной информационно-аналитической системы мониторинга и анализа кадровых потребностей рынка труда в выпускниках вузов (по номенклатуре специальностей высшего учебного заведения). Система предназначена для руководителей регионов, университетов, компаний, кадровых агентств. Развитие системы позволит руководителям регионов координировать открытие новых вузов или осуществлять перепрофилирование существующих в соответствии с актуальными экономическими задачами, руководителям вузов – корректировать учебные программы в соответствии с изменениями на рынке труда, компаниям – эффективно осуществлять подбор персонала и его подготовку, студентам – выбирать собственную траекторию обучения.

1 Введение

Общеизвестно, что система профессионального образования должна отвечать потребностям рынка труда, быстро адаптироваться к его изменениям. Но на практике осуществимость такой гибкой настройки для инертной и консервативной образовательной системы очень сложна.

В настоящее время учебные заведения активно взаимодействуют с работодателями. Типичными формами такого взаимодействия являются участие работодателей в разработке содержания программ обучения, организация производственной практики студентов на предприятии, участие сотрудников организаций и предприятий в учебном процессе и трудоустройстве выпускников. Однако не менее важным является анализ рынка труда в части потребности в определенных умениях и навыках, предъявляемых к потенциальной рабочей силе. При этом такой анализ по его сути не входит в сферу деятельности вузов. Логично предположить, что анализ рынка труда является ответственностью структур по трудоустройству и занятости, а также местных администраций. Учебные заведения, безусловно, должны учитывать данные такого рода исследований для планирования образовательной деятельности в части подготовки специалистов, востребованных в регионе, быстро адаптируясь к изменениям в экономической ситуации. Для выпускников образование будет качественным, если оно позволит успешно конкурировать на рынке труда, получить хорошо оплачиваемую работу и сделать успешную карьеру в своей профессии. Для работодателя при приеме выпускников на работу важно не соответствие уровня их подготовки требованиям федеральных государственных образовательных стандартов (ФГОС), а их профессиональная компетентность, способность адекватно действовать в производственной обстановке. Таким образом, с точки зрения работодателей и выпускников качество образования является низким, если уровень подготовки не

Труды XVIII Международной конференции DAMDID/RCDL'2016 «Аналитика и управление данными в областях с интенсивным использованием данных», Ершово, 11-14 октября 2016

соответствует требованиям производства и рынка труда.

2 Стандартизация высшего образования и профессиональной деятельности

В системе государственной стандартизации программ высшего образования происходит отказ от жесткого нормирования содержания образования в виде набора дисциплин с фиксированной трудоемкостью и осуществляется переход к рамочной регламентации структуры образовательных программ, условий их реализации и результатов освоения, представленных в форме компетенций выпускников. Вследствие этого российские организации высшего образования получают б*о*льшую свободу в формировании образовательных программ. Это позволяет вузам точнее реагировать на запросы рынка труда, выдерживать конкуренцию на российском и международном рынках образовательных услуг.

Федеральным законом № 122-ФЗ “О внесении изменений в Трудовой кодекс Российской Федерации и статьи 11 и 73 Федерального закона "Об образовании в Российской Федерации”, вступившим в силу 1 июля 2016 г., вводится обязанность применения работодателями профессиональных стандартов в части требований к квалификации, необходимой работнику для выполнения определенной трудовой функции, если эти требования установлены законодательством или нормативно-правовыми актами РФ.

Таким образом, профессиональным стандартам отводится роль согласования требований к квалификациям сферы труда и сферы образования.

3 Проблемы рынка труда и подготовки выпускников

На данный момент для российской экономики характерно несоответствие количественного и качественного состава выпускников вузов потребностям рынка труда. Низкий уровень трудоустройства выпускников связан с дисбалансом спроса и предложения на рынке труда, качеством подготовки специалистов, несоответствием компетенций выпускников требованиям работодателя, а также с различными социальными факторами. Что касается «востребованности» выпускников вузов на рынке труда, то по данным портала «career.ru» в 2014 году (к сожалению, портал не предоставляет результатов более поздних исследований) в список «Top-20» российских вузов, чьи выпускники были наиболее востребованы на рынке труда, попали два вуза из Санкт-Петербурга, остальные – из Москвы [1]. Этот факт подчеркивает остроту регионального аспекта проблемы. Анализ был проведен на основе поисковых запросов работодателей.

По данным исследовательской компании MAR Consult, изучавшей, работают ли люди по профессии, полученной в ходе обучения в вузе, по специальности не работают около половины (52%) участников исследования. Опрос проводился в Москве, Санкт-Петербурге, Екатеринбурге, Нижнем Новгороде и Самаре [2].

Проблема прогноза экономического развития и подготовки соответствующих специалистов актуальна для многих стран, в том числе европейских, где также более востребованным становится исследование потребности в квалификациях на региональном и местном уровне, а также на уровне отдельных предприятий. Анализ опыта деятельности по прогнозированию потребностей экономики в квалификациях в странах ЕС позволяет сделать вывод о том, что единых системных подходов анализа рынка труда с позиций изменений требований к квалификациям рабочей силы и отражении перспективных потребностей сферы труда в содержании образовательных программ не выработано [3].

Эффективное прогнозирование потребностей в кадрах рынка труда возможно только на основе объективной оценки состояния рынка. Научно-практический интерес к данной проблеме подтверждается разработками информационно-аналитических систем, предназначенных для автоматизации сбора данных с популярных сервисов по поиску работы, и их последующему анализу с целью выявления наиболее востребованных на текущий момент специальностей и профессий (см., например, [4]), расчета и выдачи по запросу основных индикаторов состояния рынка труда районов и региона в целом [5].

Таким образом, представляется целесообразной разработка и развитие автоматизированной информационной системы (АИС) мониторинга соответствия кадровых потребностей рынка и уровня подготовки выпускников.

4 Разработка АИС

Целью проекта является обеспечение дополнительных возможностей для выявления качественных и количественных связей между сферой образования и рынком труда. Способом достижения цели является разработка автоматизированной информационной системы мониторинга и прогноза ситуации на рынке труда и анализа кадровых потребностей по номенклатуре специальностей высших учебных заведений (на примере РЭУ им. Г.В. Плеханова). Система разрабатывается с расчетом на широкий круг пользователей и предназначена в первую очередь для руководителей регионов, университетов, компаний, кадровых агентств. Ожидается, что реализация проекта позволит теснее связать систему образования в стране и рынок труда, даст возможность руководителям вузов корректировать учебные программы, руководителям регионов – открывать новые вузы или перепрофилировать существующие в соответствии с экономическими задачами регионов, компаниям – эффективно

осуществлять подбор персонала и его подготовку. Кроме того, предполагается, что система станет полезным инструментом для молодых специалистов, только что закончивших ВУЗ, студентов старших курсов, начинающих искать работу по избранной специальности, а также студентов младших курсов, определяющихся со своей специализацией.

В качестве исходных данных в разработке используются ресурсы интернет-портала https://rabota.mail.ru/, нормативные документы: утвержденные ФГОС ВО по части направлений подготовки, реализуемых в РЭУ им.Плеханова (http://fgosvo.ru/support/49/49/17); реестр утвержденных профессиональных стандартов (http://profstandart.rosmintrud.ru). Позиция сайта rabota.mail.ru в рейтинге Alexa (http://alexa.com) определяется посещаемостью ресурса в регионе RU (38 место) и свидетельствует о его популярности. Насколько контент ресурса обеспечивает адекватное отражение состояния рынка труда в целом в части публикации вакансий и резюме является предметом отдельного исследования.

4.1 Методическое обеспечение

Реализация компетентностного подхода к подготовке выпускников вузов регламентируется ФГОС ВО, обязательными к применению всеми имеющими государственную аккредитацию вузами РФ, и предполагает формирование у студентов набора общекультурных, общепрофессиональных и профессиональных компетенций. Компетенция трактуется как 1) способность применять знания, умения, навыки и личностные качества для успешной деятельности в различных профессиональных ситуациях; 2) интегральная норма качества образования межпредметного характера.

Профессиональные компетенции систематизированы под виды деятельности. Под компетентностью понимается уровень владения совокупностью компетенций, степень готовности к применению компетенций в профессиональной деятельности. Для реализации ФГОС ВО по соответствующему направлению подготовки образовательное учреждение разрабатывает основную профессиональную образовательную программу (ОПОП), которая включает учебный план, учебный график, рабочие программы дисциплин (модулей) и практик, фонды оценочных средств, методические материалы и другие компоненты. Планируемые результаты освоения образовательной программы (набор компетенций) указываются в общей характеристике ОПОП, в рабочие программы дисциплин (модулей) включается перечень планируемых результатов обучения по дисциплине (модулю), соотнесенных с планируемыми результатами освоения образовательной программы (компетенциями). Во многих вузах ведется активная деятельность по разработке и апробации компетентностной модели выпускника (КМВ) с целью комплексно описать выпускника как субъект, обладающий готовностью применения знаний, умений, навыков и личностных качеств вести продуктивную профессиональную деятельность. В КМВ обычно входят: характеристика профессиональной деятельности, требования к результатам освоения ОПОП (перечень компетенций), таблица отношений между компетенциями и учебными дисциплинами ОПОП, паспорта компетенций (совокупность требований ВУЗа к уровню сформированности компетенции по окончании освоения ОПОП, а также развернутая характеристика требований к результатам образования в части конкретной компетенции). Это очень сложная и масштабная работа, которая еще не завершена. Таким образом, со стороны системы образования для анализа доступны формулировки содержания компетенций.

С точки зрения профессиональной деятельности можно говорить о компетентностной модели специалиста (КМС) как субъекта, востребованного на рынке труда. Эту модель описать еще сложнее, т.к. работодатели не ограничены формальными рамками формулирования текстов об имеющихся вакансиях. Как было отмечено в п.3, ожидается, что утвержденные профессиональные стандарты могут стать связующим звеном между требованиями к квалификациям сферы труда (КМС) и требованиями к результатам обучения сферы образования (КМВ).

В виду вышеизложенного на данном этапе реализации проекта за основу взяты упрощенные КМВ и КМС. Выпускник описывается как субъект, обладающий набором компетенций, сформированных во время обучения по основному виду профессиональной деятельности в рамках направления подготовки по заявленному профилю. Основу КМС составляют требования, выставляемые работодателями к рабочей силе в текстах объявлений о вакансиях.

Идея описания предметной области в виде иерархической модели, представляющей собой ориентированный граф, вершины которого соответствуют объектам предметной области, а рёбра задают отношения между ними, была заимствована из работы [6]. Построенные по такому принципу модели позволяют связать требования рынка и образовательные компетенции на различных уровнях. При этом образовательная модель вуза содержит 5 иерархических уровней («факультеты», «направления подготовки», «профиль», «вид деятельности», «содержание компетенций»), а модель рынка труда – 4 уровня («сферы деятельности», «направления», «профессии», «требования»). Отображение одной модели на другую происходит посредством установления связей на их нижних уровнях: для вуза это – «компетенции», для рынка труда – «требования» (рис. 1). Реализация связей на нижних уровнях позволяет, поднимаясь по иерархии снизу-вверх, получать связи на любом из выбранных (в зависимости от решаемой задачи) уровней. Например, отображать связи «направление подготовки» – «профессии».

Рисунок 1 Взаимное отображение между моделями системы образования и рынка труда на разных уровнях иерархий

4.2 Математическое обеспечение

Сбор и обработка данных осуществляется на основе современных методов и технологий получения информации из web-ориентированных источников. На следующем этапе применяются алгоритмы машинного обучения для перевода слов в векторное представление. После чего рассчитываются вектора предложений, что позволяет выявлять смысловое сходство требований рынка труда и профессиональных компетенций высшего образования, представляющих собой не что иное как короткие текстовые предложения. Полученные результаты используются для выявления связей на высших уровнях иерархии, описанных ранее в тексте статьи, и визуализации результатов.

4.2.1 Алгоритм сбора и обработки данных

Алгоритм сбора данных реализован в виде периодически запускаемых заданий, каждое из которых выполняет свою часть работы с данными:

1) поиск новых объявлений по ключевым словам (название должности, работодателя, регион, заработную плату, список обязанностей, список требований), которые задаются в виде параметров и позволяют ограничить предметную область;

2) сбор и загрузка объявлений в базу данных;

3) выделение из текста объявлений значимых областей (название вакансии, региона, зарплаты, требований, обязанностей);

4) подготовка текстов требований рынка труда для дальнейшего связывания.

4.2.2 Алгоритм связывания требований рынка и образовательных компетенций

Как известно, моделирование семантики (смысла) слова – одна из ключевых проблем, относящихся к обработке естественного языка (Natural Language Processing, NLP). Результаты семантического анализа используются в поисковых системах [7], системах автоматического перевода [8] и других областях, связанных с обработкой текста на естественном языке.

На текущий момент в подходах векторного представления слов (англ. Word embedding) лидирующее место занимают так называемые предсказательные модели, основанные на использовании нейронных сетей [9]. Одним из главных инструментов для векторного представления слов является word2vec [10] – группа связанных моделей, использующих нейронную сеть прямого распространения и алгоритм Continuous Bag of Words [12], предсказывающий слово по контексту, а также т.н. распределенный Skip-gram [13], предсказывающий контекст по слову. Существуют попытки создать предсказательную модель для перевода документа в векторное пространство [14]. Однако задача сравнения коротких предложений на смысловую схожесть обладает определенной спецификой и использование существующих моделей по переводу слов или документов в векторное пространство без модификаций дает неудовлетворительный результат.

Поскольку тексты формулировок образовательных компетенций, также как и формулировки требований в объявлениях о вакансиях, содержат в среднем около 10 слов, то в основе аналитической части системы лежит задача вычисления семантической близости двух коротких предложений. Авторами был разработан алгоритм перевода предложений в векторное пространство, основанный на известной нейронной модели дистрибутивной семантики – «word2vec». Данная модель, обученная на корпусе Российской Википедии и Национальном корпусе русского языка, производит отображение слова в n-мерное метрическое пространство. Таким образом, каждому слову в соответствие ставится вектор размерностью n, которая характеризует слово и влияет на точность модели. Метрическое пространство отображений слов принято называть семантическим.

В качестве примера на рис.2 представлены проекции векторов на плоскость. Близкие по смыслу

слова находятся рядом и образуют некоторые смысловые кластеры.

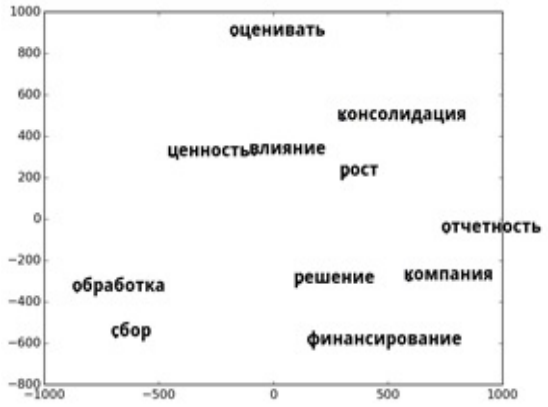


Рисунок 2 Распределение векторов слов в проекции на плоскость

Векторное представление позволяет вычислять «похожесть» слов на основе расчета косинусного расстояния. Так, для двух слов w_1 и w_2, представленных в виде векторов $\vec{V}(w_1)$ и $\vec{V}(w_2)$ семантическая близость рассчитывается по формуле:

$$cos\left(\vec{V}(w_1), \vec{V}(w_2)\right) = \frac{\vec{V}(w_1) \times \vec{V}(w_2)}{\left|\vec{V}(w_1)\right| \cdot \left|\vec{V}(w_2)\right|}$$

По аналогии с вычислением похожести слов, рассчитывается семантическая близость компетенций и требований, которые представляют собой короткие предложения, имеющие в своем составе в среднем 10 слов. Расчет вектора описанных предложений $\vec{v}(s)$, где $s = \{w_1, w_2, \dots, w_k\}$, определяется как среднее взвешенное от векторов слов, из которых оно состоит:

$$\vec{v}(s) = \frac{\sum_{i=1}^{k} p_i * \vec{v}(w_i)}{\sum_{i=1}^{k} p_i},$$

где p_i – вес слова, который рассчитывается как отношение частоты употребления слова к размерности лексикона выбранного уровня иерархии на стороне системы образования или рынка труда, а k – количество слов в предложении. После чего рассчитывается семантическая близость предложений по формуле, приведенной выше. Стоит отметить, что слова, не имеющие смысловой нагрузки (союзы, частицы, предлоги, местоимения и так далее), не участвуют в формировании вектора предложения.

Одним из методов визуализации результатов сравнения является построение взвешенного графа (рис. 3) отражающего связи между отдельными компетенциями и требованиями рынка труда (вершины соответствуют либо образовательным компетенциям, либо требованиям рынка труда и различаются цветом, а ребра – отражают наличие и силу (косинусное расстояние) связи между ними).

Вычислительная сложность алгоритма достаточно велика, и время построения матрицы смежности после применения процедуры распараллеливания на сервере с 48 ядрами (96 потоков) Intel Xeon E7-8850 v2 частотой 2,30 ГГц и 256 ГБ ОЗУ может быть оценено примерно в 4 часа.

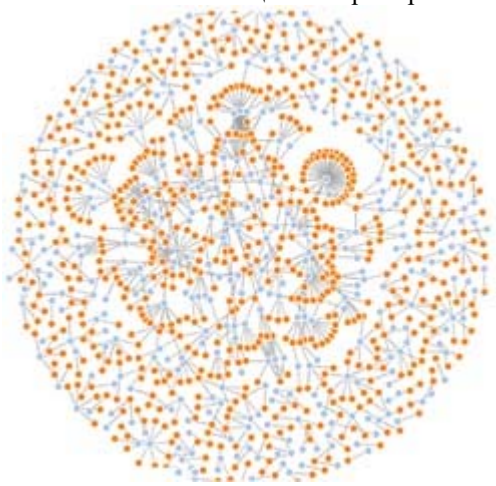


Рисунок 3 Взвешенный граф, отражающий связи между образовательными компетенциями и требованиями рынка труда, накопленными в базе данных системы (для данных, собранных на настоящий момент в базе системы).

4.2 Перспективы развития алгоритма

В силу того, что сравниваемые предложения имеют узкую направленность, а Российская Википедия и Национальный корпус русского языка охватывают огромное число сфер и видов деятельности, модель получается довольно размытой относительно задачи. Главным образом это проявляется в отсутствии векторов для некоторых слов либо их вариаций. Для частичной ликвидации подобного эффекта принято решение сделать модель двухуровневой: второй уровень представляет собой тот же алгоритм сравнения предложений, что описан выше, однако, он работает не со словами, а с основами слов, то есть с их неизменными частями. Авторы предполагают, что накопление базы вакансий позволит сформировать уникальный корпус, учитывающий специальную терминологию рынка труда, который затем будет использоваться в обучении моделей.

Также стоит отметить, что подтверждение адекватности результатов сравнения теоретически возможно с использованием вмешательства экспертов, однако объемы полученных результатов свидетельствуют о фактической невозможности полноценной проверки в разумные сроки. Поэтому авторами разрабатываются методы, которые позволят верифицировать работу данной модели.

4.3 Средства реализации

Система реализована с использованием свободно распространяемого программного обеспечения, и может быть перенесена на любую операционную систему (*Microsoft Windows, Linux* и т.д.). В качестве

языка разработки выбран *Python*, в качестве хранилища данных используется СУБД *MySQL* для хранения словарей, связей и файлового хранилища для сохранения исходных текстов документов. Система работает в распределенной облачной среде на основе программной платформы с открытым исходным кодом *OpenNebula*.

4.4 Характеристика АИС

Реализованный прототип автоматизированной информационной системы представляет собой web-ориентированное приложение с интуитивно-понятным пользовательским интерфейсом, обеспечивает надежное хранение данных.

Система построена по модульному принципу и включает:

- модуль сбора текстовых данных (функционирующий в автоматическом режиме с использованием открытых источников, в качестве которых выступает интернет-портал, аккумулирующий информацию кадровых агентств);
- модуль загрузки и хранения данных, состоящий из базы данных и распределенного хранилища данных (обеспечивающего репликацию и архивирование);
- модуль автоматической обработки, выполняющий подготовку информации для анализа, автоматическое связывание требований и компетенций, машинное обучение;
- модуль генерации и отображения отчетов;
- пользовательский интерфейс.

На рис. 4 представлен общий вид главного окна системы.

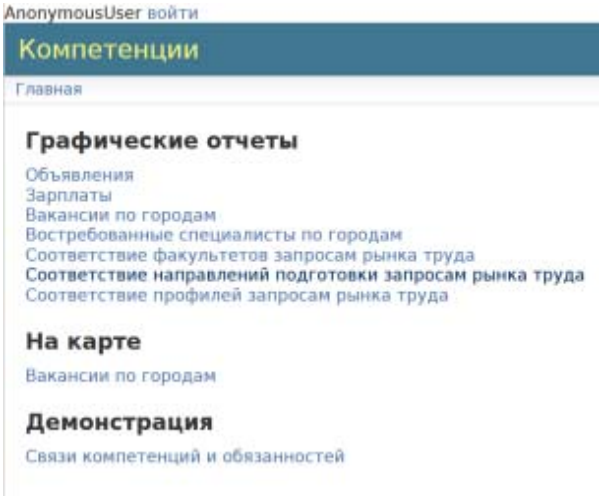


Рисунок 4 Главное окно web-интерфейса прототипа системы

Генератор отчетов позволяет просматривать их как в текстовом, так и в графическом виде (как результат предварительного анализа). Также можно проводить количественный анализ, например, выявлять наиболее популярные требования к кандидату для той или иной профессии, или выявлять профили направления подготовки вуза, соответствующие наибольшему количеству вакансий на рынке и т.д. В качестве примера работы системы на рис.5 представлен результат анализа соответствия профилей подготовки в РЭУ им. Г.В. Плеханова требованиям рынка труда.

Рисунок 5 Круговая диаграмма, отражающая «востребованность» профилей подготовки в РЭУ им. Г.В. Плеханова на рынке труда (на январь 2016 года)

Другим примером является анализ востребованности рынком труда выпускников РЭУ в городах России. Поскольку карта связанности содержит информацию о связи между компетенциями и требованиями, можно установить связь «компетенция-вакансия». Исходя из полученного распределения (рис.6), можно сделать заключение, в каких городах выпускникам РЭУ будет проще найти работу.

5 Заключение

В рамках реализации проекта создан прототип автоматизированной информационной системы мониторинга и анализа кадровых потребностей регионов РФ по номенклатуре специальностей вуза. Прототип разработан для решения задач РЭУ им. Г.В. Плеханова, в том числе – для включения в состав программных и технологических решений.

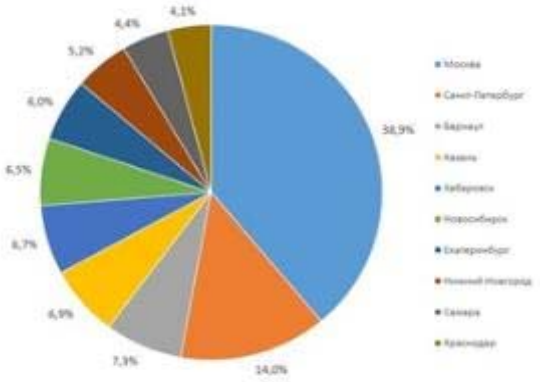


Рисунок 6 Круговая диаграмма, отражающая потенциальную востребованность выпускников РЭУ в различных городах РФ

Ситуационного центра социально-экономического развития России и регионов РФ. С помощью этой системы, в результате анализа постоянно обновляющихся больших массивов данных, можно устанавливать, насколько программы обучения высшего образования соответствуют текущим ожиданиям рынка, прогнозировать изменение этих ожиданий и автоматически выдавать рекомендации для корректировки учебных программ с целью наиболее точного соответствия этим ожиданиям. Развитие и адаптация системы могут производиться в соответствии с требованиями заказчика в зависимости от специфики задачи – особенностей региона, отдельного вуза и т.д. Созданная система, а также алгоритмы и принципы, на которых она построена, в дальнейшем могут быть использованы и для решения более широкого класса актуальных проблем. Для этого система может быть перенастроена в зависимости от особенностей в постановке задачи и типа входных данных.

* В рамках выполнения НИР «Автоматизированная информационная система мониторинга и анализа кадровых потребностей рынка (первый этап)» из средств ФГБОУ ВО «РЭУ им. Г.В. Плеханова»

Литература

[1] Материалы портала CAREER.RU.-URL: https://career.ru/article/15115.

[2] Погорелов Е. Проблема востребованности выпускника вуза на современном рынке труда// Материалы V Международной студенческой электронной научной конференции. «Студенческий научный форум» 15 февраля - 31 марта 2013 года.

[3] О.Н. Олейникова, А.А. Муравьева. Прогнозы потребности в умениях и профессиональное образование и обучение – опыт ЕС // Центр изучения проблем профессионального образования — Режим доступа: http://www.cvets.ru/Modules/SNA-EC.pdf — Загл. с экрана, 22.07.16.

[4] Е.Н. Черемисина, В.В. Белага, Ю.И. Самойленко. Информационно-образовательная среда для обучения информационным технологиям на базе Института системного анализа и управления Университета «Дубна» // «Открытое образование», 2/2014 - с.59-65

[5] Петрунина О.Е. Проектирование информационно-аналитической системы управления региональным рынком труда// Современные наукоемкие технологии. – 2005. – № 5 – с.75-78

[6] Гущин А. Н. Обеспечение учебного процесса, построенного на стандартах ФГОС-3, средствами информационных технологий // Образовательные технологии. 2013. № 4. С. 84–89.

[7] Efrati, Amir (March 15, 2012). “Google Gives Search a Refresh”. The Wall Street Journal. Retrieved July 13, 2012.

[8] Eva Martínez Garcia, Cristina España-Bonet, Lluís Màrquez (May 2015). “Document-Level Machine Translation with Word Vector Models”. Proceedings of the 18th Annual Conference of the European Association for Machine Translation (EAMT), pages 59-66.

[9] Barkan, Oren (2015). “Bayesian Neural Word Embedding”. arXiv:1603.06571

[10] Mikolov, Tomas; et al. "Efficient Estimation of Word Representations in Vector Space". arXiv:1301.3781v3 [cs.CL] 7 Sep 2013.

[11] Розенблатт, Ф. Принципы нейродинамики: Перцептроны и теория механизмов мозга. — М.: Мир, 1965. — 480 с.

[12] Harris, Zellig. "Distributional Structure". Word 10(2-3):146-162.· August 1954.

[13] Broder, Andrei Z.; Glassman, Steven C.; Manasse, Mark S.; Zweig, Geoffrey (1997). "Syntactic clustering of the web". Computer Networks and ISDN Systems 29 (8): 1157–1166. doi:10.1016/s0169-7552(97)00031-7

[14] Le, Quoc; et al. "Distributed Representations of Sentences and Documents". arXiv:1405.4053.

Monitoring of the Labour Market Needs for University Graduates Based on Data Intensive Analytics

P.V. Zrelov, V.V. Korenkov, N.A. Kutovskiy, A.S. Petrosyan, B.S. Rumiantsev, R.N. Semenov, I.A. Filozova

This paper analyzes a problem of objective estimate of the labour market and training graduates that meet the expectations of employers, it gives a description of an automated information - analytical system of monitoring and analysis of employment needs of the labour market for the graduates of higher education institutions (according to the nomenclature of specialties of higher educational institutions). The system is intended for regional authorities, universities, companies, and recruitment agencies. The development of the system will allow one to more closely link the educational system and the labour market. It would enable the governors to coordinate the launching of new universities or to make the conversion of the available ones in accordance with the current economic challenges, while the leaders of universities will be able to adjust their curricula in accordance with the changes in the labour market, the companies - to effectively implement recruitment and training, and the students - to choose their own learning way.

Database Migration Project: Bridging Industry-Academia Gap

© George Chernishev[1,2]
g.chernyshev@spbu.ru ,

© Viacheslav Galaktionov[1]
viacheslav.galaktionov@gmail.com ,

© Evgeniy Klyuchikov[1]
evgeniy.klyuchikov@gmail.com ,

© Valentin Grigorev[1]
valentin.d.grigorev@gmail.com ,

© Kirill Smirnov[1]
kirill.k.smirnov@gmail.com ,

© Andrey Terekhov[1]
a.terekhov@spbu.ru

[1]Saint-Petersburg State University, Saint-Petersburg, Russia
[2]JetBrains Research, Saint-Petersburg, Russia

Abstract

An important goal of every university is to graduate students competent enough to work in their field with minimal additional training. However, industry reports that the level of newly graduated students is insufficient. Due to various reasons, a gap exists between industrial requirements and academic education.
Many typical relational database courses can be divided into two groups: those that teach how to use a DBMS and those that teach how a DBMS actually works. We propose a practice-oriented approach to teaching database systems. We have employed an extracurricular project for developing a relational database migration tool. This approach is free from some drawbacks of both types of courses: it does not suffer from the lack of time for practice, and, at the same time, it offers a decent theoretical basis. Moreover, it provides students with a broader view of a database technology by offering an opportunity to compare two DBMSes. In this paper, we describe our experience running the project for the third-year students.

1 Introduction

Every university strives to graduate students who would be able to work in industry. This is especially true for actively developing or emerging areas of information management that are prone to workforce shortage. Big data brought forth a number of such areas, namely data science and data analytics [3, 6]. McKinsey Global Institute predicts that by 2018 the shortage of specialists in these areas will be in the range of 140000-190000 [2] in the US alone. In addition, the volume of data both stored and used by various organizations are rapidly increasing [1, 2]. Thus, the demand for such skills would only grow.

The database community has responded to these challenges by starting to analyze these requirements and their place in curricula. For example, there was an attempt to teach big data in middle school [14]. Moreover, on a panel at ACM SIGMOD/PODS 2014, the current trends in teaching data Science and classic databases were discussed [17]. The panelists advocate the idea that the relational algebra, transactions and schemas are important for data scientist as well as linear algebra. Thus, teaching topics in classic database courses are quite important to provide technology knowledge necessary for various data science and data analytics applications.

Industry is interested in keeping the training expenses as low as possible, but, today, the companies indicate that there is a gap between university training and employer requirements [7, 20, 28]. This is especially true for the information technology field [10] and for the database area in particular [21] Several universities acknowledge the problem in the IT field and have already responded [19]. Let's discuss the reasons why this gap exists in the area of database teaching.

There are a large number of approaches to teaching database-related courses, which can be classified into two types [23]: courses for application developers (user-oriented) [13] and courses for database developers (system-oriented) [9, 25].

A typical user-oriented course shallowly describes the notions of DBMS and SQL. The basic theoretical aspects like data models, simple database-related algorithms are usually included. It also instructs to design a simple database and create an application that uses it. At the same time, an advanced course describes topics such as indexing, concurrency control, buffer management and so on. Usually, it also requires students to implement some module or subsystem in either a toy DBMS [22, 23], an industrial database [9] or completely from scratch [25].

Proceedings of the XVIII International Conference «Data Analytics and Management in Data Intensive Domains» (DAMDID/RCDL'2016), Ershovo, Russia, October 11 - 14, 2016

This approach worked perfectly twenty years ago. However, novel information management fields have emerged, such as: data mining, social network analysis and recommender systems. The number of database systems and their types has also increased: column-stores, graph databases, NoSQL DBMS and in-memory systems. At the same time, today, even the classical relational DBMS can be designed in a multitude of ways due to the appearance of different niches.

The outcome of all this is that the number of ideas worth teaching has increased greatly but the amount of time has usually been left unchanged. This has led to the emergence of the following drawbacks:

- A user-oriented database course is enough to illustrate fundamentals, but not enough to provide a complete understanding of a database technology even on a user level. Most topics, including SQL, are studied quickly through several toy examples. In addition, only core features are considered in these courses, no complete description is provided. Thus, there is a high probability that a newly graduated student would require additional training.
- An advanced course also fails to provide a complete understanding of its topics. Database technology has greatly evolved in the past 45 years, so it is impossible to fit this knowledge into one or two semesters and ensure a good amount of practical assignments. Students who have completed this course may also require additional training if they want to become DBMS developers or go into database research.
- Both of these courses usually employ only one DBMS. This means that students consider DBMS features from only one point of view. Moreover, different DBMSes offer different sets of even relatively basic features, for example, partitioning by reference [16] is an Oracle-specific technology.

In order to narrow the gap we propose to make students familiar with different DBMSes, to demonstrate how they differ in basic features, and to enhance students' programming skills and database knowledge. These activities are done within an additional extracurricular programming project. Its goal is to implement a prototype of a database migration tool. This tool should automate the migration from one relational database to another. This includes transfer of schema, data, and code. Code reengineering is also to be performed, although at the later stages of the project.

This project was designed for a certain group of third-year undergraduate students. These students have already chosen the area of specialization and completed a database course which includes both basic and advanced topics.

2 Approach

The main idea of our approach is to provide students with an opportunity to gain a significant amount of practical experience while also broadening their knowledge of the underlying theory. Participating students must have some basic database- and translation-related knowledge, e.g. understand what a parser is, be familiar with SQL and so on. These prerequisites can be fulfilled by completing appropriate courses.

We believe that the database migration problem suits this purpose quite well. Even small differences between many concepts' implementations in different DBMSes would require students to revise their theoretical knowledge. Studying two DBMSes makes it possible to spot both strong and weak points of them and positively contributes to the overall understanding of DBMS technology. Students can learn about different language constructs, types of indexes and so on. Moreover, the project allows not only to master narrow understanding of two specific DBMSes, but also to see the big picture.

Our approach is designed for a small group of up to five participating students. At least two specialists – one in databases and the other in translation – should supervise the group. In order to manage the development process, students and instructors should conduct regular live meetings, where the results would be discussed and new goals would be set. In addition to that, mentors should provide online video sessions to help and motivate the students.

The project milestones can be easily designed in a way that would allow incremental improvement of both the prototype and the students' knowledge. Schema, data and code translation represent the main stages. Each of them also can be divided further to smoothly raise the level of difficulty.

Finally, the project is an interdisciplinary one. It lets students apply and consolidate the knowledge received from courses on databases, language translation and software reengineering.

3 Project

This section describes our initial experience running the project with a group of third-year students.

3.1 Organization

There are several methodologies for organizing and running student projects, e.g. [12, 28]. Let us examine the organization of our project.

Firstly, the project was described to a group of interested students and the prospects were evaluated. The students investigated possible approaches to database migration and surveyed existing tools. After that, supervisors answered all the questions and commented on the subject.

It is well known that it is not possible to fully automate the database migration process. In many cases, DBA's attention is required to resolve the situation. In this project we, like many other developers of similar migration projects, aim to automate database migration to the largest possible extent. In addition, we do not aim to develop a full-fledged industrial tool, but instead we concentrate on teaching students.

These considerations lead to the following choice: the proposed migration system is designed to operate in an

offline manner, i.e. the database is disconnected from the network and is not serving client requests.

We designed the development process to have a live meeting every two weeks. This allows maintaining a comfortable pace while minimally distracting students from the studies at the university and at the same time enables the project to advance.

In order to facilitate the process of development, students were given a Subversion[1] repository and a bug tracking system. Source and target DBMSes were installed and configured on their personal computers. Additionally, students were given the freedom to select a programming language. Eventually, C++ was chosen.

Many researchers indicated the lack of real databases employed in the teaching field [18, 27]. In our project, we addressed this issue by employing a real database, supplied by an industrial company. This database was used to test our migration tool.

Similar student projects are frequently run at the Department of Mathematics and Mechanics of Saint-Petersburg State University [8, 24].

3.2 Stages

We have designed the plan to organize the development process. Each stage, if necessary, starts with a planning phase, where all relevant discrepancies between source and target DBMSes are estimated. Such approach may not be applicable in all situations, but our main goal is to train students, which imposes several restrictions:

- Long monotonous tasks, i.e. analyzing all aspects of source and target DBMSes, had to be curtailed. This lowers the quality of teaching, but helps to keep the students interested in the project.
- During the run of the project, participating students might leave, and new students might join. So, we had to introduce checkpoints where newcomers can easily join in.

Let us consider our plan in detail:

1. Planning. The first stage of our project involved surveying of the subject area. The students collected and read the whitepapers of existing systems and analyzed them. Then, they had to select their approach to the problem, plan the core components and features, and set intermediate goals and milestones.
2. Database schema migration (tables, indices, constraints, standard types). During this stage the students learnt data types of both target and source databases and devised a mapping between them. They also indicated possible type-related conversion problems, e.g. lack of precision in different datatypes, type capacity problems and large objects (LOB) limits. Next, the students studied physical design structures which were also to be migrated: indexes, materialized views, partitioning schemes. As a result, they developed a tool that implements the mapping if possible, or logs messages for a DBA otherwise.
3. The goal of this stage was to develop script-generating code which would ensure correct data transfer. Here, the main obstacle was the various minor differences such as the famous "null vs empty string" problem [4, 5]. Performance improvements were left for the future.
4. Testing. The strength of our approach is the usage of real, industrial databases. In our project, student tested their migration tool on several real-life databases. Moreover, not only did students test their implementation on the databases, but they also learnt from them. A real-life industrial database can show various usage patterns or other important details not considered in class. Thus, our project is not limited to reading documentation, which is boring, but also offers interesting problems with real-life flavour. Assignments given in the classroom can rarely offer such opportunities.
   One important aspect of testing is ensuring the data correctness in a sense that no record from the source should be lost and no new record should appear. Since we decided to migrate inactive databases, we can employ a simple validation mechanism like comparing hash values of the whole tables.
5. Code migration, our current stage (triggers, stored procedures and so on). For the sake of simplicity, we perform one-time offline translation of database code. The translation is straightforward, complex code transformations are left for the latter stages.
   Our students familiarized themselves with the grammar of the source language and wrote a parser using a compiler-compiler. Another option was to develop a hand-coded recursive-descent parser [26]. Parsers of this kind have a very simple, clear structure and, at the same time, is powerful enough for the SQL language. However, it requires a bit more time and supervisor attention, so we decided not to use it in the test project.
   Having the parser, the students should familiarize themselves with possible intermediate representations and choose useful ones. This is where we are in the project right now. After that, they will write a code generator for the target language.
6. Code reengineering. Later, we plan to perform code reengineering by detecting predefined patterns of code and substituting them with more efficient versions that are specific to the target DBMS. There are various approaches to this problem [11, 15, 29].

Since our team is currently at the code migration stage, the last two stages may require further adjustments. It is also worth noting that we left some database migration features (e.g. complex partitioning

[1] http://subversion.apache.org/

mechanisms, view management, workflow) for the future work.

3.3 General Application Design and Used Technologies

For the sake of completeness, we will describe the used technologies and the data flow (see Figure 1) in our current application. Since we have just started the code migration stage, this part is not presented.

The students decided to implement the tool as a script generator. This decision was based on the fact that migration – as well as translation – process may require verification, modification, or both from a DBA. This is especially true at the earlier stages of the project, when the tool is quite raw.

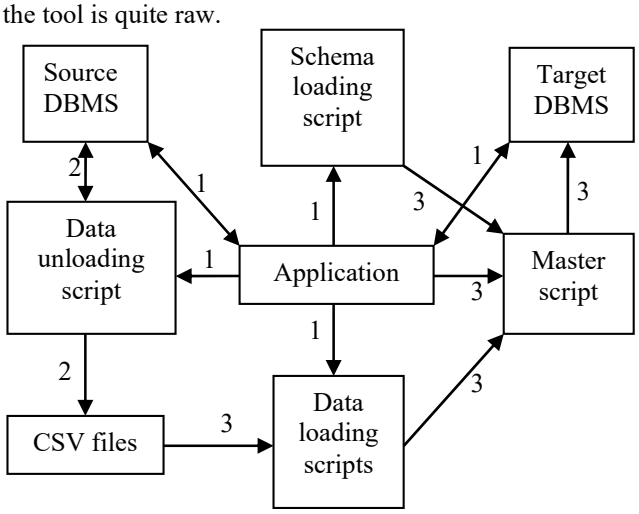


Figure 1 Application data flow

Our program is written in C++, which is popular among developers and employers. Moreover, both the source and the target DBMSes have convenient standard C++ APIs.

Each schema migration script (DDL) considers only one type of object to load (e.g. indexes, tables, etc). This may be useful to identify code that may require DBA attention or to perform data migration to a database without foreign key constraints.

The data unloading and loading scripts (DML) allow data transfer via CSV files using standard utilities such as COPY in PostgreSQL. These utilities usually provide decent performance and do not require maintenance.

Finally, the application creates a master script, which serves as an entry point to the ones already generated.

To make things more clear, let us describe the intended usage pattern of our application:

1. A DBA starts the generator, giving it access to both databases.
2. After getting the required information, the tool creates scripts for schema and data transfer.
3. The DBA then reviews these scripts in order to find the incompletely translated parts, denoted by special messages (we have mentioned them earlier).
4. If any are found, the DBA makes appropriate changes.
5. Lastly, the DBA executes the master script.

4 A case study

In this section, we present several illustrative examples confirming that students' knowledge of the field has improved. These examples are excerpts from reports prepared by students. For the most part, they are based on the comparison of two DBMSes, an unnamed commercial DBMS-X and PostgreSQL.

- Data types in PostgreSQL and DBMS-X differ greatly. Firstly, the data types of DBMS-X are under strong legacy influence, many of the data types are deprecated and overlapped. This is not the case with PostgreSQL. At the same time, PostgreSQL has a very limited support for LOB datatypes, only BLOB (bytea) exists by default. On the contrary, DBMS-X does not suffer from this kind of problem. It is also more flexible when it comes to changing the date and numeric formats and supports both national and global character sets.
- PostgreSQL has a single global namespace for all entities (indices, tables and so on), while in DBMS-X each type of entities has its own namespace.
- During the creation of a unique or a primary constraint, a unique index is created automatically.
- PostgreSQL does not have bitmap indices exposed to user; hash-indices exist but their use is officially discouraged. At the same time, DBMS-X supports bitmap indices, but hash-indices are absent. It clearly shows that different DBMSes can have feature incompatibilities, which is often unclear from the theory-oriented courses.
- Usually DBMSes have a large number of physical-layer parameters to tune. During the migration process, the students discovered that PostgreSQL has few such parameters related to index storage parameters (e.g. fillfactor, buffering). However, DBMS-X has a very complicated system of physical layout parameters. This aspect was also absent in the theoretical courses.
- Students learnt about use-cases that involve complex transactional constraints management related to DEFERRED/IMMEDIATE mode.

5 Outcomes

The main result of our project is well-trained students and methodology approbation. During the project, the students deepened their skills in a number of ways. They have acquired experience of:

1. working in a closer-to-industry environment than their fellow students who have not participated in such projects.
2. simple project management, including discussing intermediate results with accomplished industrial developers and academic researchers.

3. working with two distinct DBMSes, an experience which the majority of their fellow students lack. They have also improved their knowledge of DBMS technology.
4. participation in a practical science-intensive interdisciplinary project.

The following artifacts were created during the lifetime of the project: its source code, documentation, presentations and related materials. Let us summarize it:

1. A survey of white papers on migration tools and analysis of source and target DBMSes documentation were performed.
2. A schema migration module was implemented. Currently it aims for schemas without advanced DBMS features.
3. A data migration module, which accounts for the majority of discrepancies like those in the date and number formats, was implemented.
4. The development of a parser for the source language is close to completion now. The students already understand the grammar of both languages for the most part.

6 Drawbacks

However, our approach is not perfect. During the project run, we have been able to discern the following drawbacks:

- Our project requires relatively trained and motivated students who are willing to improve their DBMS knowledge and work hard.
- There is no straightforward way to continue this project next year. We can suggest several options:
  1. Pick up where the project was left off, concentrating on the database reengineering part. This option offers several quite interesting problems but requires working with the old team and/or advanced newcomers.
  2. Start from scratch with a new group of students; the old prototype is not used and the efforts of the old group are wasted.
  3. Employ different DBMSes and, as a result, collect a set of migration tools over the years.

  The choice heavily depends on how advanced a new group will be.
- We depend on industry, which provides us with real datasets, additional supervisors and experienced developers to consult.
- It is unclear whether our approach is better than part-time employment or summer internships. To investigate this issue we have to perform a comparison with a control group. Unfortunately, we cannot do that because our project runs for the first time and has few participants.

However, despite the aforementioned drawbacks, we believe that our project has the right to life.

Firstly, the students have indeed improved their skills. The experience of working in a realistic team project and a better understanding of DBMS field have made them more valuable for industry. The gap has been narrowed.

Secondly, our project offers a good value for money, if such costs are considered. Usually, such additional training requires extensive supervisor attention. Several student projects which we have seen [8], required about 15 hours of individual instructor time per week. Instructors' attention become one of the dominating components in the overall cost of the project. But the topic and the approach we propose lead to high autonomy of students, which in turn allows to drastically reduce instructor involvement, thus lowering the costs.

7 Conclusions

In this paper, we have described our practice-oriented approach to teaching database courses. In the introduction section, we argue that successful teaching of data science requires a solid background in database technology. Transactions and schemas are important systems-level features that cannot be easily left up to the application programmer [17]. Thus, these topics should be taught for data scientists too. At the same time, industry reports insufficient level of preparedness of newly-graduated students in the database area [21].

In our study, we concentrate on improvement of teaching classical database-related courses. Our idea is to employ an extracurricular project dedicated to the database migration problem. The goal is to construct a tool that would automate database migration. The core of this paper is the description of our experience running this project for the third-year students.

Firstly, we surveyed the existing approaches to teaching database courses and highlighted their drawbacks. Then we described the goals of our project, its organization and milestones. Next, we provided some details regarding the current status of our migration tool. To assess the value of our approach, we discussed the outcomes and drawbacks. Despite the discovered shortcomings, we believe that our approach has the right to live.

Acknowledgments

We thank anonymous DAMDID 2016 reviewers for their insightful feedback on earlier versions of this manuscript.

References

[1] Big Data: 20 Mind-Boggling Facts Everyone Must Read. http://www.forbes.com/sites/bernardmarr/2015/09/30/big-data-20-mind-boggling-facts-everyone-must-read

[2] Big data: The next frontier for innovation, competition, and productivity. http://www.mckinsey.com/business-

functions/business-technology/our-insights/big-data-the-next-frontier-for-innovation

[3] Data Scientist: The Sexiest Job of the 21st Century. http://www.tias.edu/docs/default-source/Kennisartikelen/harvard_data-scientist-the-sexiest-job-of-the-21st-century_2012.pdf?sfvrsn=0

[4] Database migration from Sybase ASE to PostgreSQL. https://wiki.postgresql.org/images/6/60/Pgconf-Sybase-to-postgres_en.pdf

[5] Empty Strings and NULL - Oracle to SQL Server Migration in PowerBuilder. http://www.sqlines.com/powerbuilder-oracle-to-sql-server/empty_string_null

[6] Insight: Data Science Fellows Program. http://insightdatascience.com/insight_white_paper.pdf

[7] Students Think They're Ready For The Real World; Employers, Not So Much. http://www.forbes.com/sites/realspin/2015/09/21/students-think-theyre-ready-for-the-real-world-employers-not-so-much

[8] А. Н. Терехов. Вспоминая о статье «Как готовить системных программистов». Компьютерные Инструменты в Образовании, (4):3–13, 2007.

[9] Ailamaki and J. M. Hellerstein. Exposing undergraduate students to database system internals. SIGMOD Rec., 32(3):18–20, Sept. 2003.

[10] Aken and M. D. Michalisin. The impact of the skills gap on the recruitment of MIS graduates. SIGMIS CPR'07, pages 105–111, 2007. ACM.

[11] M. Allamanis and C. Sutton. Mining idioms from source code. SIGSOFT FSE 2014, pages 472–483, New York, NY, USA, 2014. ACM.

[12] Andreescu, I. Intorsureanu, A. Uta, and R. Mihalca. Team work in software development student projects. CompSysTech'08, pages 80:V.18–80:1, 2008. ACM.

[13] Y. Bi and J. Beidler. Teaching database systems with web applications team projects. J. Comput. Sci. Coll., 23(3):82–88, Jan. 2008.

[14] P. S. Buffum, A. G. Martinez-Arocho, M. H. Frankosky, F. J. Rodriguez, E. N. Wiebe, and K. E. Boyer. CS Principles Goes to Middle School: Learning How to Teach "Big Data". SIGCSE'14, pages 151–156, 2014. ACM.

[15] Y. Cohen and Y. A. Feldman. Automatic high-quality reengineering of database programs by abstraction, transformation and reimplementation. ACM Trans. Softw. Eng. Methodol., 12(3):285–316, July 2003.

[16] G. Eadon, E. I. Chong, S. Shankar, A. Raghavan, J. Srinivasan, and S. Das. Supporting table partitioning by reference in Oracle. In Proceedings of the 2008 ACM SIGMOD international conference on Management of data, SIGMOD'08, pages 1111–1122, New York, NY, USA, 2008. ACM.

[17] Howe, M. J. Franklin, J. Freire, J. Frew, T. Kraska, and R. Ramakrishnan. Should we all be teaching "intro to data science" instead of "intro to databases"? In Proceedings of the 2014 ACM SIGMOD International Conference on Management of Data, SIGMOD '14, pages 917–918, New York, NY, USA, 2014. ACM.

[18] N. Jukic and P. Gray. Using real data to invigorate student learning. SIGCSE Bull., 40(2):6–10, June 2008.

[19] G. Kozieł. University education tailored to labour market expectations. EDUCON'12 IEEE, pages 1–5, April 2012.

[20] R. J. Leach, L. L. Burge, and H. N. Keeling. Can students reengineer? SIGCSE Bull., 40(1):102–106, Mar. 2008.

[21] Radermacher, G. Walia, and D. Knudson. Investigating the skill gap between graduating students and industry expectations. ICSE Companion 2014, pages 291–300, 2014. ACM.

[22] R. Ramakrishnan and J. Gehrke. Database Management Systems. McGraw-Hill, Inc., New York, NY, USA, 3rd edition, 2003.

[23] E. Sciore. SimpleDB: A simple java-based multiuser system for teaching database internals. SIGCSE Bull., 39(1):561–565, Mar. 2007.

[24] K. Smirnov and G. Chernishev. ACM SIGMOD programming contest: an opportunity to study distinguished aspects of database systems and software engineering. Computer Tools in Education journal ("Kompjuternye instrumenty v obrazovanii"), (5):25–32, 2012.

[25] Sotomayor and A. Shaw. ChiDB: Building a simple relational database system from scratch. SIGCSE'16, pages 407–412, 2016. ACM.

[26] L. Torczon and K. Cooper. Engineering A Compiler. Morgan Kaufmann Publishers Inc., San Francisco, CA, USA, 2nd edition, 2011.

[27] P. J. Wagner, E. Shoop, and J. V. Carlis. Using scientific data to teach a database systems course. SIGCSE Bull., 35(1):224–228, Jan. 2003.

[28] J. Whittington and K. Nankivell. Addressing student and industry needs through experiential learning courses to better prepare the student for real-world work experience. SIGGRAPH'08, pages 13:1–13:1, 2008. ACM.

[29] R. S. Xin, W. McLaren, P. Dantressangle, S. Schormann, S. Lightstone, and M. Schwenger. Meet DB2: Automated database migration evaluation. Proc. VLDB Endow., 3(1-2):1426–1434, Sept. 2010.

Семантический поиск и навигация

Semantic Search and Navigation

Навигация по тезаурусам и поиск в распределенных гетерогенных информационных системах

© О. Л. Жижимов © С. А. Сантеева

Институт вычислительных технологий СО РАН (ИВТ СО РАН),
Новосибирский государственный университет (НГУ)
Новосибирск
zhizhim@mail.ru saya_santeeva@mail.ru

Аннотация

Обсуждаются вопросы, связанные с построением пользовательских интерфейсов для навигации по статьям тезаурусов и рубрикаторов в гетерогенных информационных системах. Приводятся некоторые алгоритмы формирования этих интерфейсов с учетом привязки внешних информационных ресурсов к выбранным статьям тезаурусов и рубрикаторов. Основной акцент сделан на динамическую привязку внешних ресурсов на основе текстового поиска по наборам характеристических терминов. Описываются стенд для проведения исследований и результаты исследований на тестовых экспертных наборах данных.

Работа выполнена при поддержке гранта ведущих научных школ НШ-7214.2016.9.

1 Навигация по рубрикаторам и поиск в гетерогенных информационных системах

С развитием технологий построения гетерогенных распределенных информационных систем, включающих в себя множество различных баз данных с различной структурой и содержанием, актуальным становится вопрос поиска информации в базах данных с использованием онтологий, тезаурусов и классификационных схем, представленных в виде отдельных баз данных (БДОТК - базы данных онтологий, тезаурусов и классификаторов).

Существует множество различных способов построения БДОТК, организации доступа к их содержимому и реализации явных и неявных связей между БДОТК и другими гетерогенными информационными ресурсами. Многие из этих способов основаны на строгих онтологических моделях [1,2] и для практической реализации предъявляют очень жесткие требования к организации информационных систем и баз данных вплоть до полной перегрузки информации в промежуточные хранилища, функциональные свойства которых позволяют обеспечить выявление всех семантических связей между информационными объектами на основе заданных онтологических моделей. Такой подход имеет право на существование, однако остается открытым вопрос о том, как включить поиск семантически связанной информации в существующих распределенных гетерогенных информационных ресурсах, причем в случае, когда они не могут быть перегружены в специализированные хранилища.

Настоящая работа посвящена описанию способов поиска семантически связанной информации в распределенных гетерогенных информационных системах (базах данных) без использования специализированных технологий семантического поиска, основанных на моделях Semantic WEB [3-6]. Описание способов будет иллюстрироваться их реализацией в существующих программных продуктах, в частности, на примере программной платформы ZooSPACE [7], предназначенной для интеграции разнородных распределенных информационных систем, успешно функционирующей в ИВТ СО РАН на базе распределенных узлов в городах Новосибирск, Томск, Красноярск и Иркутск и объединяющей сегодня более 70 различных баз данных с общим количеством записей более 60 миллионов.

Несмотря на привлекательность перспектив использования технологий Semantic Web для поиска информации [8], реальность сталкивается с фактом, что подавляющее большинство информационных ресурсов, организованных в виде различных баз данных (реляционных, иерархических, сетевых и пр.), поддерживают прежде всего ту или иную булеву модель атрибутивного поиска информации [9], т.е. поиска, основанного на использовании метаданных и предопределенных индексов (точек доступа).

Нашу задачу можно сформулировать и так: требуется найти все записи в некотором множестве гетерогенных баз данных, которые бы соответствовали определенной онтологической

Труды XVIII Международной конференции DAMDID/RCDL'2016 «Аналитика и управление данными в областях с интенсивным использованием данных», Ершово, 11-14 октября 2016

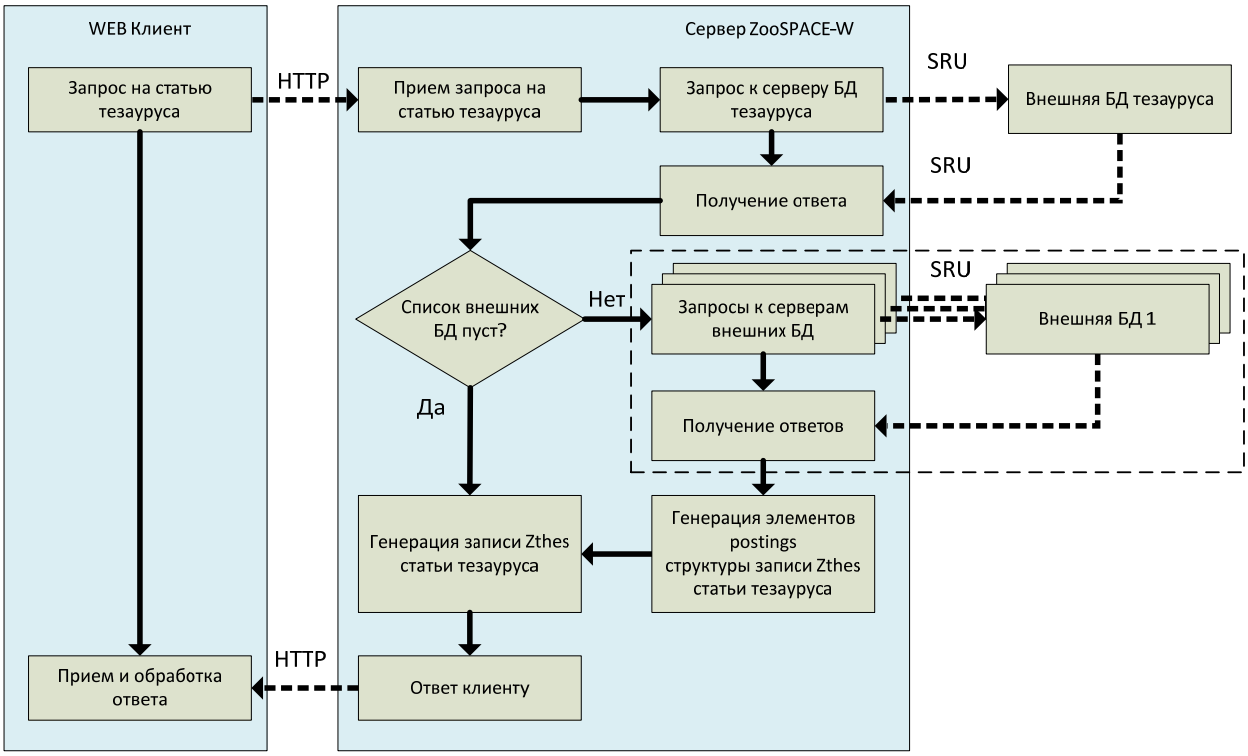


Рисунок 1 Формирование структуры записи статьи тезауруса с динамическими связями с внешними базами данных

сущности (статье тезауруса, рубрике, коду рубрикатора и пр.). Для определенности ниже эту онтологическую сущность мы будем ниже называть статьей тезауруса, понимая, что на ее месте может быть и другое. В качестве решения можно рассматривать алгоритм получения результата, реализованный в виде функционирующего серверного программного модуля для некоторой информационной системы. Эта задача практически полностью эквивалентна задачи навигации по статьям тезауруса, когда для текущей статьи тезауруса отображается информация о связанных с этой статьей записях из выбранного множества в общем случае гетерогенных баз данных. При этом «привязка» связанных записей баз данных к статье тезауруса должна быть динамической, т.е. формироваться в процессе формирования представления собственно статьи тезауруса.

Итак, клиент, используя WEB-браузер может просматривать тезаурус, перемещаясь по связанным статьям. Каждая выбранная статья тезауруса должна быть представлена клиенту в виде некоторой универсальной структуры, которая может быть однозначно интерпретирована, т.е. эта структура должна соответствовать какой-нибудь стандартной схеме данных, используемой для описания статей тезауруса. Ниже везде мы будем использовать схему данных ZThes [10] в формате XML [11]. Также мы будем подразумевать, что все необходимые обращения к серверам баз данных будут соответствовать спецификациям SRU [12] с языком запросов RPN [13] в синтаксисе PQF [14]. Этот язык запросов отличается от стандартного для SRU языка запросов CQL, но на наш взгляд он более удобен для формирования запросов и, что немаловажно, более нагляден.

На Рисунке 1 схематично представлен алгоритм работы клиента и сервера при просмотре статьи тезауруса.

Выбор клиентом статьи тезауруса порождает обращение к WEB-серверу, который в свою очередь формирует запрос к серверу баз данных, хранящему информацию о текущем тезаурусе (БД тезауруса). Этот запрос соответствует запросу на поиск записи (статьи) по ее однозначному идентификатору, в результате его выполнения должна быть получена запись БД, соответствующая требуемой статье тезауруса и содержащей полную информацию о ней.

Если клиентом был сформулирован список баз данных, записи которых следует соотнести с текущей статьей тезауруса, должен быть включен механизм формирования запросов к каждой базе данных из выбранного списка, выполнения этих запросов на соответствующих серверах БД, получение ответов и формирование специальных элементов (postings) в записи статьи тезауруса, содержащих информацию об именах баз данных и количестве найденных записей [10]. Заметим, что выполнение запросов к внешним базам данных может происходить параллельно с асинхронным завершением.

```
<term>
 . . .
  <relation>
    <termID>31.27.20</termID>
```

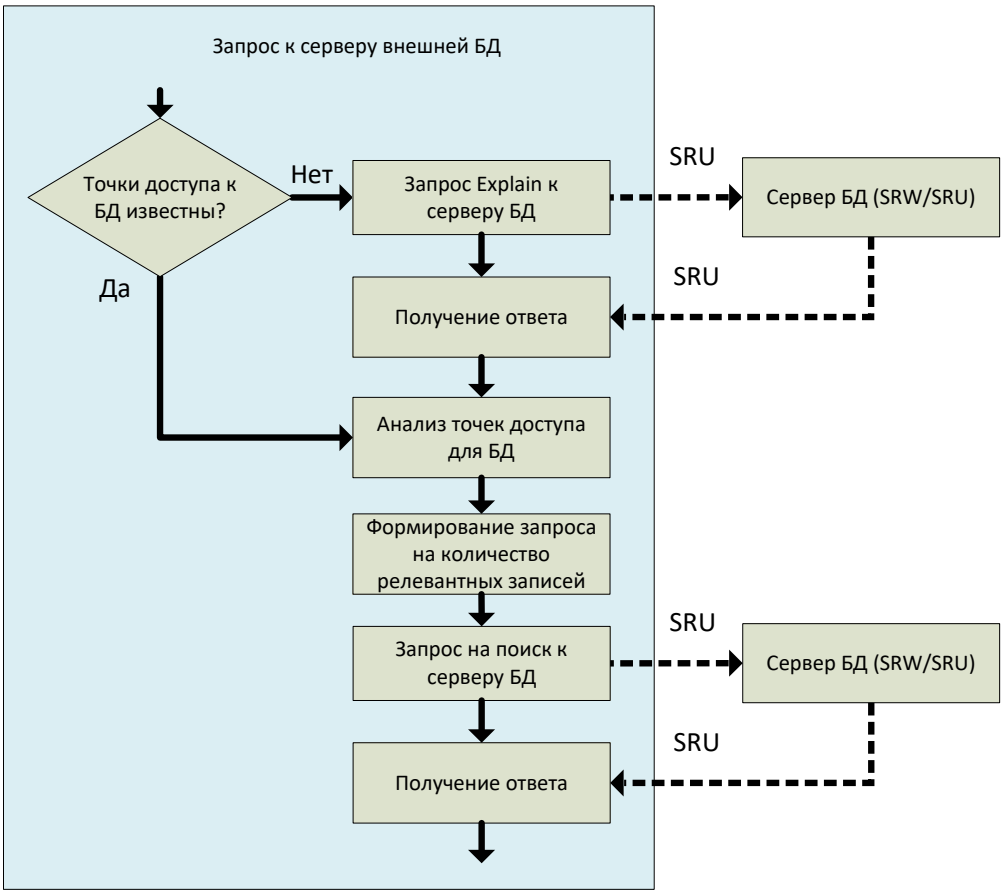


Рисунок 2 Формирование и исполнение запроса к внешней базе данных

```
    <relationType>RT</relationType>
    <termQualifier>31.27.20</termQualifier>
    <termName>Биохимия вирусов</termName>
    <termLanguage>rus</termLanguage>
  </relation>
  <postings>
    <SourceDB>AB</SourceDB>
    <hitCount>1022</hitCount>
  </postings>
</term>
```

Несомненно, самым критичным блоком этого алгоритма является блок формирования запросов к серверам внешних БД (на Рисунке 1 это блок обведен пунктиром). Именно от работы этого блока зависит качество динамической привязки записей внешних БД к текущей статье тезауруса. Работа этого блока представлена на Рисунке 2.

Прежде, чем сформировать запрос к серверу внешней БД, необходимо выяснить возможности этой БД в смысле поиска информации, т.е. в терминах SRU (или Z39.50) определить поддерживаемые поисковые атрибуты и варианты их комбинаций. Если отбросить тривиальные и маловероятные конфигурации с фиксированными точками доступа, существует только один регулярный способ – предварительно выполнить запрос explain (SRU, SRW, Z39.50) и проанализировать полученную структуру на предмет выявления поддержки требуемых поисковых атрибутов.

Например, из записи Explain можно сделать вывод, что внешняя база данных поддерживает поисковые атрибуты USE (type 1) 14 (УДК) и 21 (ключевые слова), операция сравнения - «равно» (type 2 = 3), поисковые термы интерпретируются как строки или слова (type 4=1,2,108), поиск возможен как по точному совпадению (type 5=100), так и по усечению справа (type 5=1). Поэтому к этой БД мы можем обращаться с поиском «по ключевым словам» и кодам рубрикатора УДК, т.е. если текущая статья нашего тезауруса (рубрикатора) является описанием рубрики УДК, то запрос к внешней БД должен выглядеть следующим образом (RPN в синтаксисе PQF):

```
@attr 1=14 @attr 5=1 {term}
```

где вместо «term» должен фигурировать код текущей рубрики. Следует заметить, что здесь запрос сформулирован с усечением справа, т.е. будут найдены все записи, коды УДК которых начинаются с символов «term». Для иерархических рубрикаторов это означает, что к текущей рубрике будут привязаны записи БД, содержащие коды УДК не только текущей, но и всех дочерних рубрик.

В случае тезауруса каждая статья идентифицируется ее заголовком, поэтому поиск во внешних БД следует выполнять по ключевым словам, причем по полному их совпадению:

```
@attr 1=21 {term}
```

где вместо «term» должен фигурировать заголовок текущей статьи тезауруса.

Строго говоря, такие запросы к внешним БД возможны только тогда, когда

1. Для рубрикаторов:
   a. a. для всех внешних БД возможен поиск по кодам текущего рубрикатора
2. Для тезаурусов:
   a. для всех внешних БД возможен поиск по ключевым словам
   b. ключевые слова для всех внешних БД сгенерированы из заголовков статей текущего тезауруса.

Последнее условие (2b) практически никогда не выполняется, поскольку разработчики той или иной внешней БД могут использовать тезаурусы, отличающиеся от нашего текущего, или не использовать вообще никакие, выбирая ключевые слова для записей БД в соответствии со своими правилами.

Возникает вопрос - как можно соотносить записи внешних БД с текущей статьей тезауруса при нарушении приведенных выше условий.

Для рубрикаторов при нарушении условия 1a возможны два варианта:

1. поиск по связанным кодам других рубрикаторов
2. поиск по текстовым характеристикам статьи рубрикатора

1.1 Поиск по связанным кодам других рубрикаторов

Поиск по связанным кодам других рубрикаторов может быть полезен, когда внешняя база проиндексирована по этим кодам. Действительно, если внешняя БД не проиндексирована по кодам текущего рубрикатора, например, ГРНТИ, но проиндексирована по кодам УДК, наличие связи между статьей рубрикатора ГРНТИ и статьями УДК позволяет выполнить динамическую привязку записей из внешней БД не по кодам ГРНТИ, а по кодам УДК.

```
<Zthes>
. . .
<term>
  <termID>20.23.19</termID>
  <termQualifier>20.23.19</termQualifier>
  <termName>
     Процессы информационного поиска
  </termName>
  <termType>NT</termType>
  <termLanguage>rus</termLanguage>

  <relation>
    <termID>20.23</termID>
    <relationType>BT</relationType>
    <termQualifier>20.23</termQualifier>
    <termName>Информационный поиск</termName>
    <termLanguage>rus</termLanguage>
  </relation>
  <relation>
    <SourceDB>ruudc</SourceDB>
    <relationType>RT</relationType>
    <termQualifier>025.4.03</termQualifier>
  </relation>
</term>
</Zthes>
```

Технически динамическая привязка записей из внешней БД осуществляется также, как описано выше.

1.2 Поиск по текстовым характеристикам статьи рубрикатора

Если внешняя БД не проиндексирована по кодам текущего и связанных рубрикаторов, динамическая привязка ее записей к статьям текущего рубрикатора становится задачей нетривиальной.

Действительно, для того чтобы записи из внешних БД могли быть динамически привязаны к текущей рубрике, необходимо иметь поисковый образ документов, соответствующих этой рубрике. При этом почти очевидно, что для такого поискового образа практически бесполезна текстовая информация, которая обычно присутствует в описании статьи рубрикатора (название, описание, названия связанных рубрик и т.п.). Тем не менее, можно придумать схему динамической привязки, основываясь, например, на векторной модели поиска и дополнительной информации, которой должна быть дополнена каждая статья рубрикатора.

В векторной модели поиска в качестве поискового образа выступает некоторый уникальный для каждой статьи рубрикатора вектор, определенный в многомерном пространстве в декартовой системе координат, каждая ось которой соответствуют своему уникальному термину из фиксированного списка терминов, характеризующих данную рубрику (Q_1, Q_2, ..., Q_n) [15]. Если рассматривать каждую запись внешней БД как аналогичный вектор в пространстве встречающихся в ней терминов (X_1, X_2, ..., X_m), то можно говорить о скалярном произведении векторов **Q** и **X**. Чем больше это скалярное произведение, тем выше релевантность записи **X** запросу **Q**. Критерием отбора записей может быть выполнение условия

$$\frac{1}{n}(\boldsymbol{Q}\cdot\boldsymbol{X}) \geq s,\ \ s \leq 1$$

Таким образом, для реализации динамической связи записей из внешних БД со статьей текущего рубрикатора, необходимо:

Наличие для каждой статьи рубрикатора уникального характеристического вектора. Этот вектор может быть построен только в результате обработки большого количества документов, уже имеющих в результате экспертной оценки коды рубрик текущего рубрикатора. При этом для каждой рубрики количество обработанных документов должно быть достаточно большим. Вопрос о достаточной размерности вектора, т.е. о количестве необходимых характеристических терминов зависит от структуры рубрикатора и может быть решен в результате тестов.

Определение параметра s, характеризующего минимально допустимое значение скалярного

произведения векторов при поиске может быть произведено в результате тестов.

Наличие возможности серверами БД обрабатывать поисковые запросы, соответствующие векторной модели поиска. Это требование как правило выполняется для поисковых систем, ориентированных на неструктурированную и слабоструктурированную информацию. Серверы БД ориентированы на булеву модель [9] поиска, что затрудняет использование обсуждаемой технологии привязки записей. Тем не менее в простейшем варианте без использования частот встречаемости терминов в документе и в наборе документов, поисковый запрос, соответствующий векторной модели, может быть представлен в булевом виде.

В качестве примера рассмотрим характеристический вектор длиной n=4 с терминами a, b, c, d: **Q** = (a, b, c, d)/4.

Таблица 1

	Булевый запрос (& - AND, | - OR)	**s**	**k**
1	a&b&c&d	1	0
2	(a&b&c) | (a&b&d) | (a&c&d) | (b&c&d)	0,75	1
3	(a&b) | (a&c) | (a&d) | (b&c) | (b&d) | (d&c)	0,5	2
4	a | b | c | d	0,25	3

при этом количество групп, объединенных операторами OR, равно количеству сочетаний из n элементов по k: $k!/(n-k)!$, причем $s = (n-k)/n$.

Из приведенного примера видно, что

1. При заданной длине n вектора запроса Q параметр критерия отбора a принимает дискретные значения в интервале (0<s≤1) с шагом 1/n, причем $s = (n-k)/n$.
2. Каждый булевый запрос для фиксированного s (или k) перекрывает все запросы с большими s (меньшими k).
3. При фиксированном параметре s для поиска необходимо исполнить только один запрос, который содержит $n!/k!\,(n-k)!$ групп по (n-k) термов. При этом количество участвующих в запросе термов равно $n!/k!\,(n-k-1)!$.
4. Группы объединяются оператором OR (ИЛИ), термы внутри группы объединяются оператором AND (И).

Наконец, можно сделать некоторое предположение для иерархических рубрикаторов. Если нас интересует запрос для рубрики `N.M.L`, для которой определен характеристический вектор $\mathbf{q}_{NML}$ и соответствующий частный запрос q_{NML}, то действующим запросом для рубрики `N.M.L`, будет запрос вида

$$Q_{LMN} = Q_{NM} \,\&\, q_{NML} = q_N \,\&\, q_{NM} \,\&\, q_{NML},$$

где частные запросы q_N и q_{NM} соответствуют характеристическим векторам $\mathbf{q}_N$ и $\mathbf{q}_{NM}$ для рубрик N и N.M соответственно.

Таким образом, поиск по текстовым характеристикам статьи рубрикатора возможен и может быть реализован в соответствии с упрощенной векторной моделью поиска конвертированием векторных запросов в булеву форму.

2 Экспериментальный стенд и результаты тестирования

Для проверки качества работы описанного выше механизма поиска во внешних ресурсах по текстовым характеристикам статей рубрикаторов были использованы:

2. База данных «Рубрикатор ГРНТИ», доступ к которой предоставлялся по протоколам Z39.50 и SRU в соответствии со спецификациями Zthes на платформе ZooSPACE.
3. Специализированная база данных (СБД), содержащая записи РЖ ВИНИТИ (Информатика, Автоматика, Вычислительные науки) с проставленными экспертами ВИНИТИ кодами ГРНТИ для групп кодов:
   a. 20.*.* - Информатика
   b. 28.*.* - Кибернетика
   c. 50.*.* - Автоматика и телемеханика. Вычислительная техника.

   по 200 записей для каждого кода. Для упрощения обработки эта БД была загружена в СУБД PostgreSQL с активизацией функций полнотекстового поиска в полях Title, Subject, Abstract.
4. Для числовых характеристик, описывающих качество поиска, использовались метрики [16]:

Таблица 2

	Релевантный	Не релевантный
Найдено	a	b
Не найдено	c	d

Полнота: $r = a/(a+c)$
Точность: $p = a/(a+b)$
Ошибка: $e = (b+c)/(a+b+c+d)$
F-мера: $F = 2pr/(p+r)$

Для каждой рубрики ГРНТИ в указанных выше группах на основании частоты встречаемости терминов в различных записях СБД и выполнения запроса к СБД по этому термину был определен ранжированный по убыванию F список слов из заголовков, ключевых слов и аннотаций для соответствующих записей СБД.

На основе этого списка для каждой рубрики ГРНТИ может быть построен наиболее эффективный

по вышеуказанным метрикам характеристический вектор. Мы использовали критерий максимального значения F при варьировании параметров n и s.

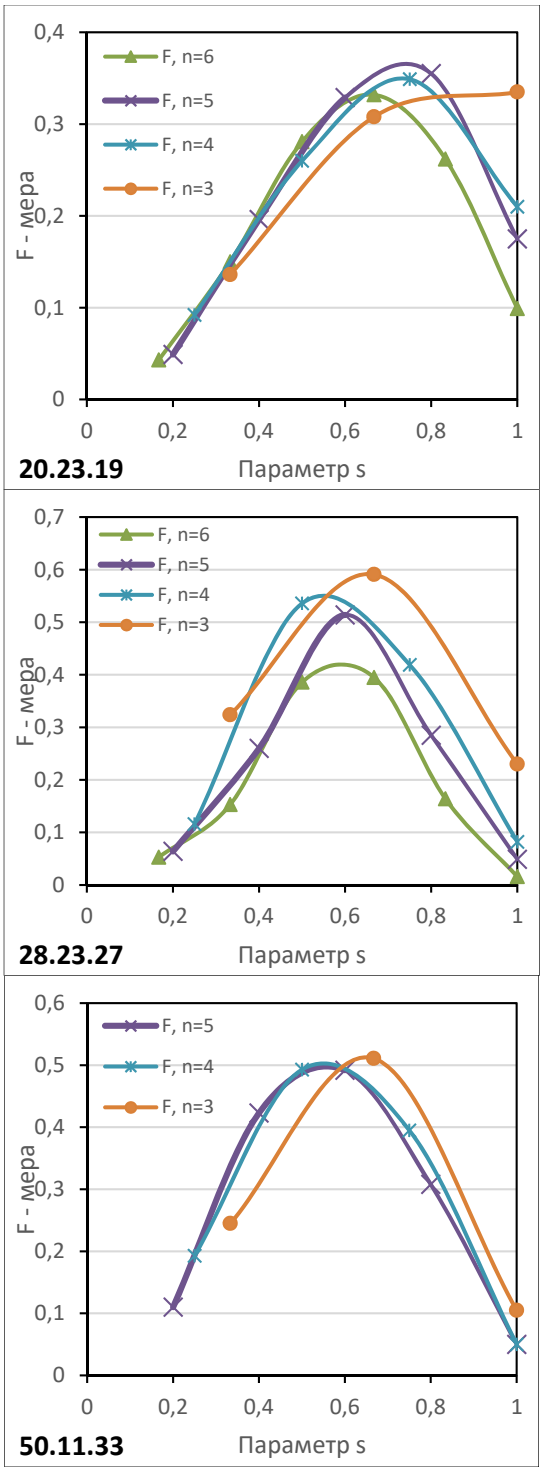


Рисунок 3 Характерная зависимость F от длины характеристического вектора (n) и параметра s для записей с разными кодами ГРНТИ.

В качестве примера приведем ранжированный список терминов для некоторых кодов ГРНТИ. Зависимость F-меры от длины характеристического вектора и параметра s представлена на рисунке 3 для трех кодов ГРНТИ. При этом для наиболее оптимальных значений n и s в таблице 4 приведены значения метрик.

Таблица 3

ГРНТИ	Термины
20.23.19 - Процессы информационного поиска	поисковый, запрос, поиск, документ, информационный, пользователь, обработка, база, …
28.23.27 - Интеллектуальные робототехнические системы	робот, мобильный, движение, алгоритм, управление, предлагаться, …
50.11.33 - Оптические запоминающие устройства	оптический, дисковод, воспроизведение, носитель, диск, запись, память, …

Таблица 4

	20.23.19	**28.23.27**	**50.11.33**
N	5	3	3
S	0,8	0,667	0,667
K	1	1	1
R	0.360	0,508	0,430
P	**0,350**	**0,707**	**0,629**
E	0,009	0,009	0,005
F	0,355	0,591	0,511

Таким образом, при наличии СБД можно определить для каждой рубрики:

1. Ранжированный список терминов
2. Длину и содержание характеристического вектора
3. Оптимальное значение параметра s (или k)

На основании этой информации можно построить булевый поисковый запрос по текстовым атрибутам, который наиболее полно будет соответствовать запросу по соответствующему коду рубрикатора. При этом вероятность нахождения нужных записей в найденном таким образом множестве записей предварительно известна и равна значению p.

Например, вместо запроса по коду ГРНТИ 28.23.27 можно выполнять запрос вида

```
(робот & мобильный)
| (робот & движение)
| (мобильный & движение)
```

Результат выполнения этого запроса будет содержать нужные данные с вероятностью 0,7.

Следует заметить:

1. Описанный механизм привязки внешних ресурсов к кодам рубрикаторов хорошо работает для «грубых» рубрикаторов.
2. Для иерархических рубрикаторов и рубрикаторов с «родственными» рубриками качество поиска является удовлетворительным. При этом поисковые метрики сильно зависят от длины

характеристических векторов и значения критерия отбора. Обе этих характеристики могут быть получены на основе анализа экспертных данных.

В заключение следует заметить, что на основе изложенных выше методов и алгоритмов в настоящее время разрабатываются программные модули для системы ZooSPACE, реализующие графические пользовательские интерфейсы для навигации по тезаурусам и рубрикаторам с привязкой информации из разнородных источников.

Литература

[1] Web Ontology Language (OWL), http://www.w3.org/2004/OWL/

[2] Онтология (информатика) — Материал из Википедии - свободной энциклопедии — http://ru.wikipedia.org/wiki/Онтология_(информатика)

[3] Semantic Web, http://www.w3.org/2001/sw/

[4] Tim Berners-Lee, James Hendler and Ora Lassila, The Semantic Web, http://sciam.com/article.cfm?articleID=00048144-10D2-1C70-84A9809EC588EF21

[5] Metadata Architecture, http://www.w3.org/DesignIssues/Metadata

[6] W3C standards, http://w3.org/sw/

[7] Жижимов О.Л., Федотов А.М., Шокин Ю.И. Технологическая платформа массовой интеграции гетерогенных данных // Вестник Новосибирского государственного университета. Серия: Информационные технологии. - 2013. - Т.11. - № 1. - С.24-41. - ISSN 1818-7900.

[8] Guha R. Semantic search / R. Guha, R. McCool, E. Miller // Proceedings of the 12th international conference on World Wide Web. – N.Y. ACM Press, 2003. – P. 700–709.

[9] Шарапов Р.В., Шарапова Е.В., Саратсвцева О.А. Модели информационного поиска. http://vuz.exponenta.ru/PDF/FOTO/kaz/Articles/sharapov1.pdf

[10] The Zthes specifications for thesaurus representation, access and navigation - http://zthes.z3950.org/

[11] Extensible Markup Language (XML), http://www.w3.org/XML/

[12] SRU - Search/Retrieve via URL// The Library of Congress. - USA - http://www.loc.gov/standards/sru/

[13] RPN - https://www.loc.gov/z3950/agency/markup/09.html

[14] Mike Taylor. PQF - http://search.cpan.org/dist/Net-Z3950-PQF/lib/Net/Z3950/PQF.pm

[15] Э. Мбайкоджи, А.А. Драль, И.В. Соченков. Метод автоматической классификации коротких текстовых сообщений. http://elib.ict.nsc.ru/jspui/bitstream/ICT/1396/1/93_102.pdf

[16] М. Агеев, И. Кураленок, И.Некрестьянов. Официальные метрики РОМИП 2006 http://romip.ru/romip2006/appendix_a_metrics.pdf

Thesaurus navigation and search in the distributed heterogeneous information systems

Oleg L. Zhizhimov, Saya A. Santeeva

The issues related to creation of the user interfaces for navigation through the articles of thesauruses and rubricators in heterogeneous information systems are discussed. The algorithms of formation of these interfaces taking into account a binding of external information resources to the chosen articles of thesauruses and rubricators are given. The main emphasis is placed on a dynamic binding of external resources based on text search in the sets of characteristic terms. The workbench for studies as well as the research results obtained on the testing expert data sets are described.

Семантический поиск как средство взаимодействия с электронной библиотекой

© Д. А. Малахов © Ю. А. Сидоренко © О. М. Атаева © В. А. Серебряков

МГУ им. М.В. Ломоносова ВЦ им. А.А. Дородницына РАН

Москва

79155155577@ya.ru sidorenkoyury@gmail.com oli@ultimeta.ru

serebr@ultimeta.ru

Аннотация

Данная работа описывает решение проблемы семантического поиска по текстам документов. В качестве примера рассматривается семантический поиск по текстам книг цифровой библиотеки LibMeta. Представлен алгоритм построения иерархии ключевых слов и кластеров путем итеративного выполнения кластеризации и выделения ключевых слов. Построенная иерархия используется для генерации рефератов и индексации документов для семантического поиска.

1 Введение

Традиционно предполагается, что ресурсы электронных библиотек представляют собой библиографические записи традиционных библиотек и электронные копии документов, описываемых этими записями. Но развитие технологий переопределяет понятие как самих библиотек, так и ее ресурсов, которые не ограничиваются только библиографическими записями и их электронным представлением, но также выводит на передний план семантику этих ресурсов. Для этого могут использоваться различные виды классификации ресурсов библиотеки. Разработаны различные отраслевые рубрикаторы, которые позволяют более детально определить тематическую направленность ресурсов. Как правило, этих средств для описания семантики недостаточно, либо со временем появляются новые требования к описанию ресурсов библиотек, что приводит как к усложнению самих описаний, так и требует значительных трудозатрат на внедрение новых способов описаний, соответствующих текущим потребностям.

Используя новые возможности, которые появляются с развитием технологий, пользователь библиотеки может использовать больше средств для работы с ресурсами цифровых библиотек, имея возможность описывать область своих интересов в терминах предметной области на основе стандартов с привлечением тезаурусов словарей и онтологий. Это позволяет ему организовывать и описывать как собственные коллекции, так и собственные ресурсы, при необходимости детализировать описания ресурсов и свою область интересов, уточняя ее термины.

Персональная открытая семантическая цифровая библиотека LibMeta[1] характеризуется гибким хранилищем метаданных для своих ресурсов и типами описываемых информационных ресурсов. Такой подход к описанию ресурсов библиотеки обеспечивает универсальность описания ее типов ресурсов и объектов независимо от предметной области и области интересов пользователей. Структурированность описания обеспечивает поддержку связей между различными типами ресурсов.

Гибкость описания ресурсов обеспечивается использованием OWL онтологий для хранения метаданных. Такой подход дает ряд преимуществ:

- возможность выполнения SPARQL запросов;
- получение дополнительных знаний с помощью логического вывода;
- упрощение интеграции с другими библиотеками;
- возможность изменения схемы под изменившиеся потребности.

Семантический поиск – поиск документов по их содержанию. Библиотека LibMeta позволяет осуществлять семантический поиск по метаданным с помощью SPARQL запросов. При этом в библиотеке не реализован семантический поиск по текстам книг.

Целью работы является улучшение качества услуг, оказываемых библиотекой LibMeta, с помощью семантического поиска по текстам книг библиотеки.

Таким образом, необходимо реализовать систему семантического поиска по текстам книг библиотеки LibMeta. Поисковая система должна находить по

Труды XVIII Международной конференции DAMDID/RCDL'2016 «Аналитика и управление данными в областях с интенсивным использованием данных», Ершово, 11-14 октября 2016

поисковому запросу на естественном языке релевантные этому запросу тексты книг с учетом семантики. Подразумевается, что для поддержки семантики будут использованы словари синонимов и гипонимов.

2 Организация семантического поиска

Существуют разные подходы к организации семантического поиска по текстам. В последние годы наиболее популярным стало семантическое аннотирование текста. Существуют различные способы решения задачи семантического аннотирования. В каждом из них документу или части документа приписывается некоторый набор семантически близких документу меток. В дальнейшем можно искать документы по этим меткам. Кроме того, можно искать документы обычным полнотекстовым поиском, а потом учитывать эти метки при работе с документом, получая больше информации с помощью них [2]. Обычно в качестве меток используются персоны, места, организации или другие субъекты [3].

Для описания меток часто используются RDF хранилища, содержащие набор понятий и отношения между ними. Некоторые методы используют информацию из Wikipedia, как из масштабного источника знаний[4]. В последнее время методы семантического аннотирования все чаще обращаются к использованию массивного, взаимосвязанного облака Linked Open Data [3, 5]. Например, с помощью средства семантического аннотирования GATE был проаннотирован Национальный Архив Великобритании (42 TB) [6].

Семантическое аннотирование не единственный способ организации поиска. Существуют решения, основанные на улучшении классического полнотекстового поиска расширением запроса синонимами. Так была создана онтология, основанная на терминах статей с помощью УДК[7], в дальнейшем она использовалась для расширения запроса пользователя. Кроме того, подход, использующий информацию о синтаксисе, морфологии и пунктуации, также кажется интересным [8]. К сожалению, описанные подходы не были внедрены и не используются повсеместно.

Было проведено множество экспериментов по использованию словарей синонимов и гипонимов для улучшения качества полнотекстового поиска. Известно, что при использовании синонимов и гипонимов растет полнота и часто существенно падает точность поиска.[9]

Особенность предлагаемого подхода в том, что индексируется не весь текст, а только его значимые части, в зависимости от задачи, это могут быть абзацы, предложения, словосочетания или проставленная человеком метка, например, хэштег. За счет изменения размера значимой части можно контролировать точность и полноту. Например, если полнота маленькая и индексируются предложения, можно попробовать индексировать сочетания предложений. Кроме того, в предлагаемом подходе не используется транзитивность синонимов и гипонимов, для каждого слова нужно явно указать слова, которые могут быть использованы вместо него, это также упрощает контроль над качеством поиска.

3 Семантический поиск на базе S-тегов

Рассмотрим модель S-тег, которая предлагается для использования при реализации семантического поиска.

Определение. Алфавитом будем называть любое конечное непустое множество. Элементы этого множества называются символами данного алфавита.

Пример. В качестве алфавита может выступать любой алфавит естественного языка.

Пусть задан некоторый алфавит A.

Определение. Термином алфавита A будем называть любой упорядоченный конечный непустой набор символов алфавита A.

Пример. Слова и словосочетания выбранного алфавита естественного языка являются терминами этого алфавита.

Пусть задано множество терминов T алфавита A

Определение. S-тегом на множестве T будем называть любое непустое подмножество T.

Пример. Поисковый запрос, представляющий собой конъюнкцию слов и словосочетаний, образованных алфавитом естественного языка A является S-тегом на множестве T, где множество T является множеством слов и словосочетаний естественного языка алфавита A.

Пусть задано множество S-тегов ST.

Пусть $\forall t \in T$ задано множество $\mathrm{THS}_t \subset T$.

Определение. Сужениями термина $t \in T$ будем называть множество:

$$R_t = \{t\} \cup \mathrm{THS}_t.$$

Пример. Если в качестве терминов рассматривать слова и словосочетания, то в качестве множества THS_t рассмотрим множество синонимов и гипонимов термина t. Тогда множество R_t представляет собой множество, состоящее из термина t, его синонимов и гипонимов.

Определение. Классом термина $t \in T$ будем называть множество:

$$Class_t = \{ st \in ST \mid st \cap R_t \neq \emptyset \}.$$

Пример. Если S-тег является поисковым запросом, как было показано ранее, тогда $Class_t$ является множеством поисковых запросов, которые включают термин t или его синонимы, гипонимы.

Определение. Сужениями S-тега $st \in ST$ будем называть множество:

$$R_{st} = \bigcap_{t \in st} Class_t.$$

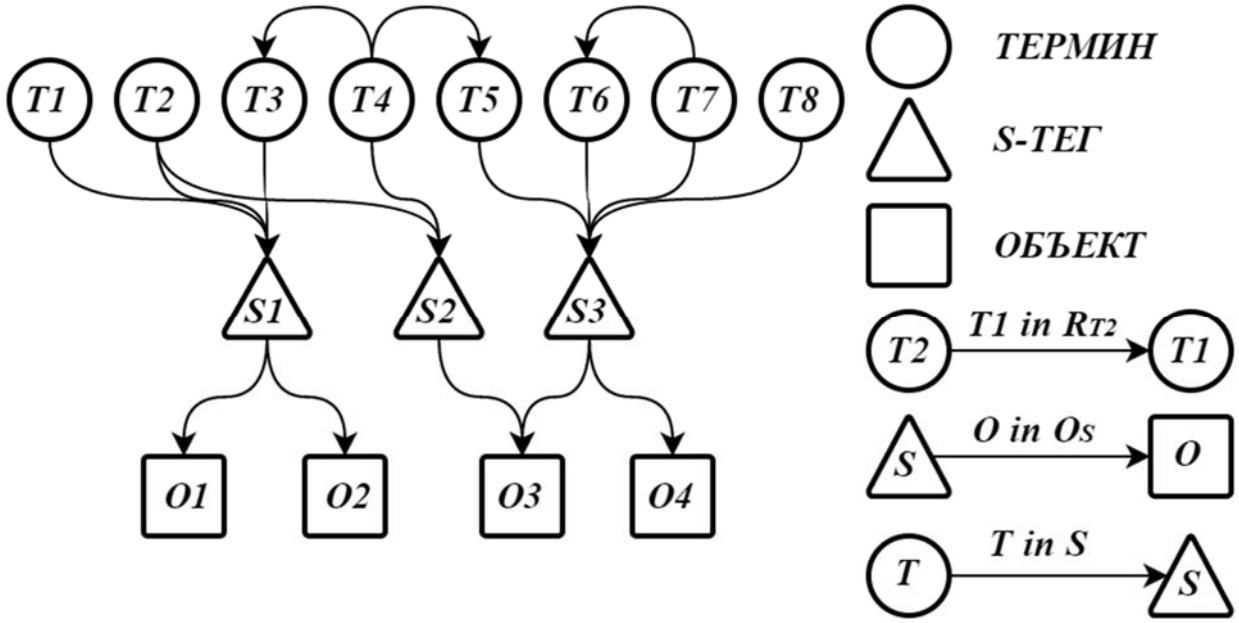


Рисунок 1 Отношения терминов, объектов и S-тегов

Определение. Классом S-тега st будем называть множество:

$$Class_{st} = \bigcap_{rst \in R_{st}} O_{rst}.$$

Пример. В качестве примера множества объектов O можно рассмотреть тексты книг. Из текста книги могут быть выделены запросы, которым этот текст является релевантным. Рассмотрим поисковый запрос st. Множество O_{st} является множеством текстов, в которых выделен запрос st. Если текст является релевантным более частному запросу по сравнению с st, то он должен быть релевантным запросу st. Отсюда следует, что $Class_{st}$ является множеством релевантных запросу st текстов книг.

На Рис. 1 представлена визуализация связей между терминами, S-тегами и объектами. Явно указано, что объект O3 ∈ $Class_{S2}$. Кроме того, объекты O1 ∈ $Class_{S2}$ и O2 ∈ $Class_{S2}$, так как S1 ∈ R_{S2}. Объект O4 ∉ Class_S2, так как S3 ∉ R_S2. Это следует из того, что термин T2 и термины из R_{T2} не включены в S-тег S3.

Под семантический поиском на базе S-тегов будем понимать поиск текстов книг, которые являются релевантными заданному поисковому запросу, который является S-тегом. Поиск может считаться семантическим, так как использует синонимы и гипонимы, позволяющие передать смысл текста.

Согласно приведенным примерам, задача поиска релевантных текстов книг по заданному поисковому запросу на естественном языке сводится к задаче нахождения $Class_{st}$ для заданного S-тега st. Рассмотрим решение этой задачи.

В первую очередь нужно найти R_{st}. Для этого достаточно найти $Class_t$ для каждого термина S-тега st.

Первый способ определения $Class_t$ требует хранения инвертированного индекса II_{ST}, где каждому t ∈ T соответствует инвертированный список S-тегов IL_t: {st | t ∈ st}. В этом случае поисковый запрос st должен быть обогащен для каждого своего термина t терминами из R_t:

$$Class_t = \bigcap_{rt \in R_t} IL_{rt}.$$

Предложенный способ требует дополнительных затрат на получение IL_{rt}. В случае большого тезауруса эти затраты могут быть значительными.

Второй способ определения $Class_t$ требует хранения инвертированного индекса II_{ST}, где каждому t ∈ T соответствует инвертированный список тегов $IL_t = Class_t$. В этом случае размер II_{ST} существенно больше, но скорость поиска выше.

Для решения задачи поиска $Class_{st}$ необходимо иметь инвертированный индекс II_O, где каждому st ∈ ST соответствует инвертированный список объектов $IL_{st} = O_{st}$:

$$Class_{st} = \bigcap_{rst \in R_{st}} IL_{rst}.$$

Решив задачи нахождения R_{st} и $Class_{st}$ для S-тега st, мы получаем решение задачи семантического поиск текстов книг, которые являются релевантными заданному поисковому запросу, как S-тегу.

4 Выделение S-тегов из текстов

Как было показано ранее, можно организовать семантический поиск по S-тегам, если они выделены из текстов книг. Рассмотрим способ автоматического выделения S-тегов из текста.

Под ключевыми словами текста мы будем понимать слова и словосочетания, которые передают смысл текста и выделяют его среди других текстов в коллекции.

Чтобы соответствовать содержанию текста, S-тег должен содержать его ключевые слова. S-тег может

являться ключевым словом, предложением или абзацем, в котором встретилось ключевое слово. Таким образом, задача выделения S-тегов может быть сведена к поиску ключевых слов в тексте.

Так как ключевое слово текста должно выделять его среди других текстов, то ключевые слова зависят как от текста, так и от коллекции текстов, которой противопоставляется этот текст. Если разбить множество текстов на группы похожих по смыслу текстов, то можно рассматривать ключевое слово как отличительный признак для текста, характеризующий его группу.

С другой стороны, если для разбиения текстов использовать в качестве признаков ключевые слова, то можно существенно повысить скорость и качество разбиения сокращением признакового пространства и фильтрацией шума.

Таким образом, кластеризация, как процесс разбиения коллекции текстов на группы, может использовать выделенные ключевые слова, в то время как алгоритм выделения ключевых слов может использовать результаты кластеризации. С помощью кластеризации можно разбить множество текстов на группы (кластеры), после чего находить ключевые слова для текстов относительно полученных групп. Этот процесс можно повторять несколько раз, чередуя выделение ключевых слов и кластеризацию.

В итоге получим иерархическую структуру документов в коллекции и соответствующую ей иерархию ключевых слов.

5 Выделение ключевых слов

Дано множество текстов D на множестве терминов T. Под термином будем понимать слово или словосочетание. Множество D разбито на множество кластеров C. Нужно выделить в текстах из множества D такие ключевые слова, которые характерны для кластера этого текста.

Традиционно выделение ключевых слов делится на два этапа. Первый этап представляет собой выделение кандидатов в ключевые слова. На этом этапе удаляются стоп-слова, могут фильтроваться части речи или, например, фильтроваться кандидаты, которые не содержатся в заголовках статей из Wikipedia. Второй этап заключается в проверке кандидатов на семантическую близость к данному тексту. Для решения этой задачи используют алгоритмы машинного обучения как с учителем, таки без него [10].

Основной особенностью нашей задачи в отличие от стандартной задачи выделения ключевых слов является то, что ключевые слова должны зависеть не только от самого документа, но и документов близких к нему с точки зрения некоторой предметной области. Предлагаемый подход к решению задач может быть улучшен с помощью алгоритмов решения стандартной задачи [10].

Рассмотрим простой подход к решению задачи. Пусть выбран случайный текст $d \in D$. Для каждого кластера $c \in C$ и термина $t \in T$. Оценим вероятность того, что $d \in c$ при условии, что $t \in d$:

$$P(d \in c \mid t \in d) = \frac{|(t \in d \wedge d \in c)|}{|(t \in d)|},$$

где $|(t \in d \wedge d \in c)|$ - количество документов из кластера c, в которых встречается термин t; $|(t \in d)|$ - количество документов из кластера c, в которых встречается термин t.

Пусть задан некоторый порог N, тогда будем считать, что термин t характеризует кластер c, если:

$$P(d \in c \mid t \in d) > N * \max_{c_i \in C}(P(d \in c_i \mid t \in d)).$$

Таким образом, для каждого документа выделяются все термины, которые характеризуют кластер этого документа и включены в этот документ, если оценка $P(d \in c \mid t \in d)$ достаточно велика.

6 Кластеризация

Дано множество документов D на множестве ключевых слов K. Нужно определить наилучшее число кластеров, на которые можно разбить множество документов D и произвести разбиение.

Для решения задачи воспользуемся методом кластеризации k-means++[11]. Он позволяет за линейное время разбить множество документов на k кластеров.

Критерием качества разбиения с параметром k будем считать значение Q_k, равное сумме среднеквадратичных отклонений центров, полученных кластеров за N итераций. Таким образом, чем меньше Q_k, тем более устойчивым и качественным является разбиение.

Если k выбрано слишком большим или слишком маленьким, то это скажется на качестве дальнейшего выделения ключевых слов. Поэтому подбирая параметр k, нужно задать верхнюю оценку k1 и нижнюю оценку k2.

Наилучшее значение k находится перебором от k1 до k2. Выбирается такое значение k, при котором значение Q_k за N итераций является минимальным.

Пример. Допустим k1 = 8, k2 =16, n =10. Тогда за 80 итераций может быть найдено наилучшее разбиение с точки зрения устойчивости с минимальным значением Q_k.

Рисунок 2 Схема взаимодействия

7 Реферирование

Реализовав семантический поиск по текстам книг, мы столкнемся с проблемой отображения результатов поиска. Можно использовать полный текст книги, удовлетворяющий запросу, аннотацию книги или заранее автоматически изготовленный реферат. Более предпочтительным кажется вариант генерации реферата книги по запросу пользователя.

Реферирование – процесс построения краткого содержания (реферата) документа. Реферирование используется для визуализации результатов поиска. Рефераты бывают статические и динамические.

Статические рефераты используются для предоставления краткой информации обо всем документе. Статический реферат формируется один раз и не зависит от поисковой потребности пользователя.

Динамические рефераты генерируются в момент выполнения поискового запроса пользователя и представляют краткую информацию о релевантных запросу частях текста.

В рамках работы был выбран простой алгоритм генерации рефератов, заключающийся в объединении всех выделенных S-тегов и их контекста в случае статических рефератов. В случае генерации динамических рефератов по запросу находятся его сужения и объединяются вместе со своим контекстом.

Предложенный алгоритм может быть улучшен с помощью алгоритма, основанного на доминантах[12]. Подобные подходы крайне популярны.

8 Архитектура системы

На Рис. 2 продемонстрирована схема взаимодействия внутри системы семантического поиска по библиотечным данным.

8.1 Загрузка данных

Данные необходимо получать из двух источников.

- Источник метаданных поставляет библиографические записи, Метаданные попадают в RDF хранилище, откуда пользователь может их получать с помощью SPARQL запросов. RDF хранилище реализовано на базе библиотеки Jena.
- Источник документов поставляет тексты книг, которые сохраняются в файловую систему. И в индексы СУБД Postgres.

8.2 Получение иерархии и ключевых слов

После поступления новых текстов запускается процесс кластеризации, а затем процесс выделения ключевых слов.

Далее для каждого кластера запускается процесс кластеризации на множестве ключевых слов, после чего для каждого кластера снова выделяются ключевые слова.

Эти два процесса выполняются поочередно, пока не будут получены иерархия текстов и множество ключевых слов.

Результаты сохраняются в СУБД Postgres.

8.3 Формирование индексов и рефератов

В качестве S-тегов используются предложения, содержащие ключевые слова.

Выделенные S-теги индексируются в СУБД Postgres. Поиск по тегам осуществляется с помощью GIST и GIN. Для каждого S-тега формируем список документов, к которым этот S-тег привязан.

С помощью выделенных S-тегов и их контекста производится генерации статических рефератов.

Для генерации динамического реферата по запросу пользователя находятся все его сужения с помощью полнотекстового поиска СУБД Postgres. На основе контекста найденных S-тегов формируется реферат.

8.4 Поиск и тезаурус

Пользователь задает запрос на естественном языке. Перед выполнением поисковый запрос обогащается множеством синонимов и гипонимов из тезауруса.

По запросу система находит его сужения, я для них списки привязанных документов. Для каждого документа формируется динамический реферат.

Пользователь может находить документы с помощью SPARQL запросов по RDF хранилищу.

Тезаурус хранится в файле, где каждому слову соответствует строка, содержащая список его синонимов и гипонимов. Редактируя файл, пользователь может влиять на результаты поиска.

Заключение

В рамках данной работы был реализован прототип системы семантического поиска по библиографическим данным и текстам книг.

На примере семантической библиотеки LibMeta была продемонстрирована актуальность данной работы. Внедрение описанных подходов позволяет улучшить качество предоставляемых библиотекой услуг.

Были рассмотрены различные подходы к реализации семантического поиска по текстам. Введена модель S-тега и задача поиска сужений S-тега. Задача поиска текста по поисковому запросу была сведена к задаче поиска сужений S-тега. Был представлен алгоритм решения задачи поиска сужений S-тега. Рассмотренная модель была использована при реализации поиска на базе СУБД Postgres.

В рамках работы рассмотрен алгоритм выделения S-тегов из текста. Для работы алгоритма необходимо множество выделенных ключевых слов.

Продемонстрирован процесс построения иерархии ключевых слов с помощью итеративного процесса сменяющих друг друга кластеризации и выделения ключевых слов. Предложенный алгоритм выделения ключевых слов позволяет использовать информацию о кластере документа. Для кластеризации был выбран алгоритм k-means++.

В качестве визуализации результатов семантического поиска по текстам был представлен подход к выделению статических и динамических рефератов.

Предлагаемые алгоритмы могут быть улучшены с помощью существующих решений, но в рамках прототипа были намеренно использованы простые решения. В рамках дальнейшей работы планируется:

- Улучшение качества предложенных алгоритмов выделения ключевых слов, генерации рефератов.
- Проведение экспериментов по улучшению качества кластеризации.
- Реализация эффективного хранилища S-тегов.
- Реализация распределенного выделения ключевых слов на Hadoop кластере;
- Переход к распределенной системе поиска;
- Проведение экспериментов по выделению S-тегов с помощью иерархии классификатора УДК;
- Использование контекстов терминов для семантического поиска;
- Реализация эффективного хранилища S-тегов.

Литература

[1] Атаева О. М., Серебряков В. А. Персональная цифровая библиотека Libmeta как среда интеграции связанных открытых данных. Труды 16-й Всероссийской научной конференции «Электронные библиотеки: перспективные методы и технологии, электронные коллекции» — RCDL-2014, Дубна, Россия, 2014 .

[2] GiannopoulosG. etal. GoNTogle: a tool for semantic annotation and search .The Semantic Web: Research and Applications, p. 376-380, Springer Berlin Heidelberg, 2010.

[3] Bontcheva K., Tablan V., Cunningham H. Semantic search over documents and ontologies. Bridging Between Information Retrieval and Databases, Springer Berlin Heidelberg, 2014.

[4] Berlanga R., Nebot V., Pérez M. Tailored semantic annotation for semantic search .Web Semantics: Science, Services and Agents on the World Wide Web, p. 69-81, 2015.

[5] Alahmari F., Magee L. Linked Data and Entity Search: A Brief History and Some Ways Ahead. Proceedings of the 3rd Australasian Web Conference, 2015.

[6] Maynard D., Greenwood M. A. Large Scale Semantic Annotation, Indexing and Search at The National Archives. Lrec, p. 3487-3494, 2012.

[7] И.В. Захарова. Об одном подходе к реализации семантического поиска документов в электронных библиотеках. Вестник Уфимского государственного авиационного технического

университета, 2009. http://cyberleninka.ru/article/n/ob-odnom-podhode-k-realizatsii-semanticheskogo-poiska-dokumentov-v-elektronnyh-bibliotekah

[8] А.Л. Воскресенский, Г.К. Хахалин. Средства семантического поиска. Международная конференция по компьютерной лингвистике и интеллектуальным технологиям «Диалог», 2006. http://www.dialog-21.ru/digests/dialog2006/materials/html/Voskresenskij.htm

[9] Н. В. Лукашевич. Тезаурусы в задачах информационного поиска. 2010

[10] Hasan K. S., Ng V. Automatic Keyphrase Extraction: A Survey of the State of the Art //ACL (1), 2014, p 1262-1273. http://acl2014.org/acl2014/P14-1/pdf/P14-1119.pdf

[11] k-means++: The Advantages of Careful Seeding. 2006. http://ilpubs.stanford.edu:8090/778/1/2006-13.pdf

[12] О.Г. Чанышев. Ассоциативные поля доминант и анализ текста. Институт Математики им. С.Л. Соболева СО РАН, 2011. http://elib.ict.nsc.ru/jspui/bitstream/ICT/1376/1/3OHT2.pdf

Semantic search as a means of interaction with the digital library

Dmitriy A. Malakhov, Yury A. Sidorenko, Olga M. Ataeva, Vladimir A. Serebriakov

This work is devoted to solving the problem of semantic search for document texts. As an example, we consider the semantic search for text of LibMeta digital library books. The proposed approach provides a hierarchy of documents keyword by iteratively performing clustering and selection of keywords. The hierarchy of documents keyword is used to generate abstracts and indexing documents for semantic search.

Электронный архив газет: Web-публикация, ассоциация информации с базой данных, создание полнотекстового поиска

© А. Г. Марчук © С. В. Лештаев
Институт систем информатики СО РАН,
Новосибирск
mag@iis.nsk.su svles@iis.nsk.su

Аннотация

В докладе описана разработанная система представления (публикации) информации архива сканированных газет на сайте. В практическом плане, в электронный архив газет входит ряд нетривиальных программных частей: отображение изображений сканов высокого разрешения на веб-сайте; распознавание текста страниц, создание полнотекстового индекса для поиска по ключевым словам, организация связи опубликованных выпусков газет и базы данных фотоархива СО РАН.

Благодарности

Авторы выражают благодарность принимающим участие в работе над проектом «Открытый архив СО РАН» сотрудникам Института систем информатики СО РАН: А.А.Фурсенко, И.Ю.Павловской, И.А.Крайневой, В.Э.Филиппову, П.А.Марчуку. Работа выполнена при поддержке гранта РФФИ 14-07-00386А.

1 Введение

Исследуется задача представления (публикации) набора данных. Исходным набором данных для проекта является множество из 24532 сканированных страниц и разворотов газеты «За науку в Сибири» (современное название «Наука в Сибири»). Необходимое требование решения - максимально удобный доступ к данным через браузер.

Аналогичную, хотя и более масштабную задачу решала команда Google в проекте Google newspapers https://news.google.com/newspapers. В другом проекте, на сайте https://Issuu.com отображаются журналы, оцифрованные и опубликованные в формате PDF.

Первично, в таких системах, как и в нашей, решается задача доступа через Интернет и Web-браузер к сканам высокого разрешения. Мы остановили свой выбор на технологии Deep Zoom [10] фирмы Microsoft. Привлекательным свойством технологии является возможность лёгкой подстройки визуального качества изображения пользователем в процессе чтения, вплоть до предельных характеристик его детальности.

Другими задачами, решавшимися в проекте, являлись: обеспечение связи сканированных изображений с базой данных и реализация текстового поиска по набору задаваемых слов.

Связь между сканированными образами и базой данных осуществляется в обе стороны, т.е. из базы данных мы имеем ссылки на конкретные места в изображениях и, наоборот, некоторые образы статей размечены дополнительной информацией и ссылками, ведущими в базу данных. Созданный в рамках проекта технологический комплекс позволяет в удобном виде выделять информационные фрагменты из изображений страниц и связывать фрагменты с элементами базы данных. Авторы не нашли подобных решений в других разработках, связанными с публикацией сканов газет и журналов.

Для того, чтобы сделать поиск по ключевым словам в текстах выпусков, необходимо распознать текст и сделать полнотекстовый индекс. В какой-то мере, Google в своей системе решает эту задачу. А Issuu.com не распознаёт текст с изображений, если предоставленный пользователем PDF состоит из них, но если PDF содержит текст, то поиск работает.

Частично результаты исследования этой задачи были опубликованы в статье [7] и магистерской диссертации [6]. В этой статье приводится краткое описание новых решённых частей продолжающегося проекта.

Труды XVIII Международной конференции DAMDID/RCDL'2016 «Аналитика и управление данными в областях с интенсивным использованием данных», Ершово, 11-14 октября 2016

Рисунок 1 Изображения страниц одного выпуска состыкованы по порядку в горизонтальный ряд и выровнены по высоте.

2 Отображение выпуска газеты

В разных системах Web-публикации сканированных копий газет и журналов, используется небольшой набор вариантов отображения страниц на графику браузера. Обычно используются традиционные имиджи и HTML. Некоторые используют публикацию через PDF формат файла (сайт Issue.com). Проблемой таких подходов является необходимость соблюдать компромисс между качеством изображений и объёмом JPEG или PDF-файлов. Для более качественной публикации «картинок» страниц, используют специализированные технологии и системы, в частности, применяют технологию Deep Zoom, для браузера доступную в рамках Silverlight. Google в этой задаче, использует свою технологию отображения карт.

В нашем решении используется Deep Zoom, для некоторых целей планируется добавить варианты PDF с текстом и технологию Open Seadragon. Последнее связано с тем, что Silverlight не набрал популярность настолько, чтобы использоваться повсеместно, в некоторых браузерах требуется разблокировать в настройках запуск Silverlight-компонента на сайтах, так как по умолчанию он блокируется для безопасности. Open Seadragon реализует ту же технологию средствами HTML-5, хотя и с рядом ограничений.

Deep Zoom и Open Seadragon в качестве источника используют «пирамидальную» группу сжатых копий изображения (DZI), каждая копия разделена на части по 256x256 пикселей в формате JPG. Это решение создаёт много служебных файлов: в 24532 сканах набора газет за 37 лет всего более четырёх миллионов JPG файлов.

Это, в свою очередь, породило проблему, которая выражается напр. в том, что копирование полного набора файлов кассеты может длиться часами. Для хранения такого объёма маленьких файлов и, что важно – оперативной выдачи, был разработан и применён формат быстрого архива без сжатия Sarc (Simple Archive). Он состоит из трёх блоков: 1) тело архива содержит без изменений

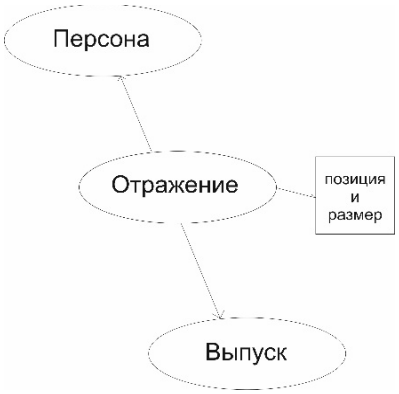


Рисунок 2 Орграф показывающий связывание в RDF данных представителей классов (название класса указано в овале) выпуска, персоны и отражения. Название текстового свойства позиции и размера области отражения указано в квадрате. Допускается множество отражений в одном выпуске, но у отражения может быть только один персонаж.

байты файлов в архиве подряд по порядку добавления в архив; 2) содержание архива в XML формате, описывающее пары: имя файла и позиция начала его байт в теле архива; 3) размер содержания в байтах, этот блок всегда фиксированного размера 8 байт (длинное целое). Объектное представление архива состоит из содержания в виде словаря пар.

Отображение выпуска осуществляется в виде состыкованных изображений страниц подходящего размера. Deep Zoom предоставляет также функциональность отображения коллекции изображений, принимая в качестве источника файлы Deep Zoom Collection, состоящие из XML описания каждого изображения коллекции: относительные размеры, позиции и путь к DZI.

Для публикации всего множества выпусков необходимо описать их в базе данных. Авторы использовали RDF СУБД, созданную на основе технологии Polar [8]. Каждая страница выпуска газеты является отдельным документом и описывается в данных как сущность класса документа. Весь выпуск описывается сущностью класса многостраничный документ и класса коллекции.

3 Взаимосвязь публикуемых сканов газет с базой данных

Привязка сканированного материала к базе данных осуществляется через отождествление области, на которой имеется изображение статьи с

Рисунок 3 Орграф примера RDF данных, показывающий связи между представителями классов выпуска газеты; отражения статьи; персонажей. Допускается множество отражений персонажей в одной статье.

документом. Далее, следует использование ассоциированных с документом отношений, таких как отражение, авторство, место.

Например, если в статье (на изображении газетного разворота) отражён или описан человек (организация или событие), информация о котором есть или может быть внесена в базу данных, информационным специалистом делается такая привязка. Такая технология опробована и используется в проекте «Фотоархив СО РАН», редактору предоставлена функциональность отметить область отражения или описания и указать персонажа. В RDF данных это записывается узлом класса отражение. Для каждого отражения указаны дуга к узлу персонажа, дуга к многостраничному документу, позиция и размер области отражения. На сайте фотоархива СО РАН на странице описания человека (организации или события) автоматически размещается ссылка на отражение, которая приводит на страницу выпуска и показывает область отражения.

Статья в выпуске рассматривается как самостоятельный информационный объект, она имеет название и прочие атрибуты описания. В RDF она записывается сущностью класса article, а множество персонажей, описанных в ней записано сущностью класса отражения. Сама статья отражена в выпуске так же, сущностью класса отражение с указанием позиции и размера.

4 Распознавание текстов

Для распознавания текста, имеющихся на сканированных изображениях авторы выбрали OCR движок tesseract, он бесплатный и формирует результирующую информацию в виде удобного формата HOCR. Это HTML со специальной идентификацией и классификацией элементов, в формате фиксируется позиция и размер абзацев, строк и каждого слова. Для этого формата можно создать конвертер в PDF. Из платных приложений ABBY Finereader распознаёт более точно, чем tesseract. Чтобы с помощью него определить позицию слова, можно распознавать изображение в формат PDF с параметром «точная копия».

4.1 Кодирование слов русского языка

Для создания базы данных всех распознанных слов, создаётся база широкого набора, в идеале – всех слов русского языка и применяется кодирование встреченных при обработке текстов слов и имён собственных. Используется список из 4 159 394 словоформ для 142 792 лемм [12] опубликованный на сайте http://www.speakrus.ru/dict/. Для кодирования с помощью технологии Polar авторами реализована таблица имён, осуществляющая биективное отображение строк в числа и обратно [4]. При этом каждое слово получает числовой код, и в каждом падеже или времени это слово имеет такой же уникальный код, т.е. слова переводятся в нормальную форму. Неинформативные слова: предлоги, союзы и т.п. называются «стоп-словами», получают специальный код.

4.2 Преобразование HOCR в поток слов

В HOCR текст уже разделён на слова, но запятые, точки и другие знаки препинания оставляются вместе со словом. Разделённые на две части слова для переноса на новую строку с помощью тире остаются двумя словами, при этом тире остаётся в конце первой части. И из-за неточности распознавания слова могут быть разделены на части. Поэтому, если слово не найдено в базе всех слов, выполняется попытка найти его конкатенацию со следующим или предыдущим словом. Кавычки могут играть роль выделения наименования, оно может состоять из нескольких слов. Для решения перечисленных проблем подходит применение синтаксического и семантического анализатора, созданного по специальной грамматике. Это исследование продолжается. Сокращения слов можно расшифровать с помощью онлайн сервиса http://www.sokr.ru/. И необходим список имён,

Рисунок 4 Изображение первой страницы первого выпуска газеты “За науку в Сибири”. На изображении выделена цветом (при печати изображение преобразовано в чёрно белое) фотография президента АН СССР Келдыш Мстислава. Под выделением размещена ссылка на страницу сайта фотоархива с его информационным портретом.

фамилий, названий. Результаты исследования и разработок применимы не только для распознанных газетных сканов, но и к другим вариантам работы с текстами.

4.3 Полнотекстовый индекс с координатами каждого слова

Всего на сканированной странице ~1500 слов, следовательно, всего на 24532 страницах менее 100 миллионов слов, что позволяет рассматривать решения размещения базы данных в оперативной памяти. После распознавания с помощью технологии Polar создаётся таблица слов, строки в ней соответствуют словам (всем, кроме стоп-слов) в страницах газет и содержат код слова (4 байта), целочисленный идентификатор страницы (4 байта) из базы данных страниц и позицию слова (4 байта x, 4 байта y). Всего не более 100 миллионов строк, не более 1,6 гигабайт на таблицу, она помещается в оперативную память. Это является достаточным условием быстрой работы с ней.

5 Пользовательский поиск

Пользователю предоставляется функция поиска [5], он указывает несколько ключевых слов, или наименований, возможно с опечатками или в каком-то падеже или времени. Для определения искомых ключевых слов используется нечёткий поиск [3], [11]. Пока есть ограничения на работу с именами людей: фамилия, имя, отчество должны совпадать полностью (кроме падежа) и в образце, и в тексте, в качестве частного послабления, в тексте возможно указано только фамилии и имени или фамилии с инициалом имени. Названия организаций часто состоят из нескольких слов или аббревиатуры, при поиске требуется совпадение этих слов в такой же последовательности.

В качестве результата поиска пользователю предоставляется множество выпусков в виде списка ссылок, отсортированных по убыванию релевантности найденных в них страниц. В одном выпуске может быть найдено несколько страниц, для каждой в списке содержится отдельная ссылка. При переходе по ссылке отображается соответствующая ей страница, и во всём выпуске подсвечиваются точками позиции всех искомых слов.

Оказалось, что предоставить пользователю найденные сканированные страницы недостаточно – ему трудно увидеть эти слова в образе страницы и появляется ощущение, что эта страница предоставлена ошибочно. В настоящее время делается попытка решить эту проблему в двух направлениях. Если пользователю по ссылке показать PDF вариант страницы, то он может самостоятельно найти на ней искомое слово с помощью поиска браузера. Другой вариант – найти искомые слова внутри демонстрируемого образа показать пользователю копию, в которой найденные слова окрашены.

5.1 Нечёткий поиск

Для того, чтобы найти множество похожих слов на данное слово не используя перебор всех слов применяется нечёткий поиск. Один из вариантов нечёткого поиска выполняется с помощью триграмм

[1], то есть троек символов, был реализован авторами. Предварительно, каждое слово в базе слов преобразуется в неупорядоченное множество триграмм, т.е. буквенных троек, «вырезанных» из слова с помощью скользящего окна, например: слово → { __с, _сл, сло, лов, ово, во_, о__ }. Каждому трёхбуквенному коду сопоставляется множество слов, в которых она содержится. При поиске искомое слово тоже преобразуется в множество триграмм.

Нечёткость поиска означает, что данное для поиска слово может отличаться от искомого заменой одного символа, пропущенным, лишним символом, перестановкой символов или соседние символы переставлены местами.

Сравним множество триграмм двух похожих слов: 1) если в одном слове символ заменён другим, то множества отличаются ровно тремя триграммами; 2) если в одном из слов вставлен лишний символ, то его множество содержит триграмм на одну больше и множества отличаются двумя триграммами; 3) если в одном из слов перестановка соседних символов, то множества отличаются на 4 триграммы.

Следовательно, чтобы найти похожие слова с точностью до одной из перечисленных ошибок, для каждой триграммы определяется список содержащих её слов. Можно перебрать все варианты выборок 4х из списков (без повторов). Для каждого варианта вычислить пересечение остальных списков (без выбранных четырёх). Объединение этих пересечений является результатом. При объединении подсчитывается количество копий каждого слова, чем больше, тем больше триграмм совпало, тем больше точность совпадения.

Если введённое слово не найдено, то нечёткий поиск найдёт список возможных вариантов искомого слова, отсортированный по убыванию точности совпадения. Для Яндекса, Google и т.п., возможно, лучше выбрать самое популярное из вариантов, используя статистику поисковых запросов и частоту появления слов в текстах [12]. В реализованном авторами поиске используются все варианты. К каждому найденному искомому слову добавляется множество его синонимов. Список синонимов русского языка можно найти в интернете [2]. Синонимы далее считаются одинаковыми словами.

5.2 Расчёт релевантности страниц

Чем больше в тексте страницы найдено различных искомых слов, тем больше её релевантность. Когда на нескольких страницах одинаковое число различных искомых слов, то релевантность можно различить по длине наименьшему из расстояний между различными словами.

6 Генерация отражений информации фотоархива

Связывание массива обработанных страниц газеты «За науку в Сибири» производилось с базой данных, вручную сформированной при обработке фотографий и списков из истории Сибирского отделения РАН, так называемым фотоархивом СО РАН [9].

Полнотекстовый поиск позволяет выполнить автоматическое выявление страниц, на которых есть упоминание имён персон, организаций, событий и др. (далее объекта) из базы данных фотоархива и зафиксировать это соответствие. Как результат, на изображениях страниц некоторым точкам или областям устанавливается гиперссылка на описание, имеющееся на сайте фотоархива.

Для того, чтобы информационный оператор проверил правильность ссылок, идентификаторы всех страниц, на которых найдено хотя бы одно наименование сохраняются в специальный список. После проверки и возможной коррекции, список уничтожается.

7 Исправление ошибок распознавания

Полиграфия газет обычно не слишком высокого качества. Поэтому большое количество символов распознаются неправильно. Планируется создать приложение для их исправления. Для этого используется база всех слов. Множество всех слов из всех распознанных страниц, которых нет в базе слов объединяются в один список. Далее для каждого ошибочного слова из списка автоматически подбирается список правильных вариантов из базы слов, это те, которые отличаются одним символом (или вставленным/удалённым символом). Вероятность того, что при распознавании символы поменяются местами мала. Более вероятны различия в 6 триграммах, когда два символа в слове неправильные. Для приложения планируется создать пользовательский интерфейс, позволяющий просматривать контекст и в один клик заменять слово на его верный вариант, или записать слово в наименования.

Заключение

Основными тезисами доклада являются: публикация сканов выпуска с помощью технологии Deep Zoom и формата архива Sarc; взаимосвязь отражённых на страницах выпуска объектов и объектов базы данных; распознавание текста с изображений и организация пользовательского поиска.

Описанные результаты в основном доведены до программного решения, применённого при обработке газеты «Наука в Сибири». В силу тематической близости, удобным оказалось

устанавливать связь с базой данных фотоархива СО РАН, результат размещён сайте фотоархива http://soran1957.ru. Опубликовано ~1700 выпусков газеты «За науку в Сибири» с 1965 по 1997. Редакторы разметили области, отражающие значимые персонажи и организации. Ранее проведено распознавание текста из сканов этого архива с помощью OCR движка Cunei, но позиции слов не вычислены. По распознаванию с помощью движка tesseract проведены успешные эксперименты на единичных страницах, требуется применение для массового распознавания всех сканов архива. Проведены эксперименты создания полнотекстового индексирования (без позиций), создана база слов, и проверен нечёткий поиск. Завершены успешные эксперименты отображения страниц (отдельно от выпуска) с помощью Deep Zoom, Open Seadragon, создания текстовых PDF (которые могут индексировать поисковики). Остальные исследования и разработки продолжаются.

Литература

[1] Ukkoncn, E. Approximate string-matching with q-grams and maximal matches. Theoretical Computer Science 92 (1992) 191-211. [Электронный ресурс] – Режим доступа: https://www.cs.helsinki.fi/u/ukkonen/TCS92.pdf (дата обращения: 29.05.2016).

[2] Абрамов. Н. Словарь синонимов. Полная парадигма. Морфология. Перевод в текст Александр Ильин, 2003. http://www.speakrus.ru/dict/

[3] Васильева О. В. Методы поиска и представления информации о расписании вуза. диплом. работа. Новосибирск. НГУ, 2015.

[4] Карасюк П.К. Технологии создания и использования больших таблиц имён. диплом. работа. Новосибирск. НГУ, 2015.

[5] Кристофер Д. Маннинг, Введение в информационный поиск / Кристофер Д. Маннинг, Прабхакар Рагхаван, Хайнрих Шютце, перевод Д. Клюшин, 2014.

[6] Лештаев С.В. Архитектура и программное обеспечение архивных фактографических систем: работа с многостраничными растровыми изображениями. дис. … магистра. Новосибирск. НГУ, 2012.

[7] Лештаев С.В., Марчук А.Г. Система создания электронных архивов газет с поиском по ключевым словам, // Системная информатика. — 2014. — № 3. — С. 1-11.

[8] Марчук А.Г. "PolarDB – система создания специализированных NoSQL баз данных и СУБД" // Моделирование и анализ информационных систем. Т. 21, № 6 (2014), с.169–175.

[9] Научная статья про фотоархив [Электронный ресурс] – Режим доступа: http://soran1957.ru/ (дата обращения: 29.05.2016).

[10] Официальная страница Deep Zoom [Электронный ресурс] – Режим доступа: https://www.microsoft.com/SilverLight/deep-zoom/ (дата обращения: 29.05.2016).

[11] Сметанин Н. Нечёткий поиск в тексте и словаре. 9 марта 2011 [Электронный ресурс] – Режим доступа: https://habrahabr.ru/post/114997/ (дата обращения: 25.05.2016).

[12] Хаген М. Частотный словарь. Полная парадигма. Морфология. Совмещённый словарь. 2014. http://www.speakrus.ru/dict/

Digital newspaper archive: Web-publication, linking with database and creating full-text search

Alexander G. Marchuk, Sergey V. Leshtaev

The paper describes the system for creating, maintaining, and Web publishing of scanned set of newspaper pages. The properties of this system include: the use of Deep Zoom technology for visualization of images of pages, text recognition and building the search index, providing of links between scanned pages and database objects.

Technology described was used to publish newspaper “Za nauku v Sibiri” and integration of newspaper articles with the database of the archive soran1957.ru

Анализ паттернов в рекомендательных системах

Pattern Analysis in Recommender Systems

Кластеризация профилей пользователей в рекомендательных системах поддержки жизнеобеспечения на основе реальных неявных данных

© С. А. Филиппов © В. Н. Захаров © С. А. Ступников
© Д. Ю. Ковалев
Институт проблем информатики ФИЦ ИУ РАН,
Москва
stanislav@philippov.ru VZakharov@ipiran.ru ssa@ipi.ac.ru
dm.kovalev@gmail.com

Аннотация

Данная работа посвящена описанию решения ключевой задачи в контексте построения рекомендательных систем поддержки жизнеобеспечения. Этой задачей является выявление пользовательских предпочтений и формализация их посредством формирования поведенческих профилей с последующим выявлением групп пользователей со схожими характеристиками. Основным источником информации о пользовательских предпочтениях является массив неявно собираемых данных об их действиях при навигации по страницам Интернет магазинов. Под поддержкой жизнеобеспечения понимается круг задач по обеспечению населения необходимыми для их жизнедеятельности продуктами, включая продукты питания, бытовой химии, косметики и многое другое. Эти задачи, как правило, решают магазины (в том числе и интернет-магазины) оптовой и розничной торговли. Авторами работы представлен подход к построению рекомендательной системы, основанный на решении задачи коллаборативной фильтрации с использованием методов кластерного анализа данных для выявления групп пользователей со схожими предпочтениями. Достоинства решения продемонстрированы на примере тестового массива данных, полученного из действующего интернет-магазина Thaisoap. Работа выполнена при поддержке Министерства образования и науки РФ, уникальный идентификатор проекта RFMEFI60414X0139.

Введение

Современная электронная коммерция, сопровождающая и поддерживающая процессы жизнеобеспечения, активно использует рекомендательные системы для решения задач адресного продвижения товаров и услуг с учетом конкретных пользовательских предпочтений. Основным источником информации о пользовательских предпочтениях являются данные об активности пользователей при посещении конкретного интернет-ресурса. Эти данные собираются в основном неявным образом (протоколирование действий пользователей) и обладают следующими основными свойствами: значительный объем и быстрое изменение (или обновление) данных во времени. При этом адаптация под конкретного пользователя – весьма сложная задача, поскольку для ее решения необходимо принимать во внимание как присущие человеку неопределенность и спонтанность в рамках конкретного интернет-ресурса, так и множество неопределенностей, связанных с особенностями функционирования Интернет.

Одним из простейших подходов к выработке рекомендаций является использование статистических метрик для выявления, например, наиболее популярных, дешевых (дорогих), близких по заданным характеристикам объектов и предложение их пользователям без учета их персональных предпочтений. Более сложные алгоритмы выявляют предпочтения пользователей посредством формирования поведенческого профиля пользователя, который, в свою очередь, определяется на основании анализа его активности при выборе товаров и услуг. Также в настоящее время развиваются подходы, основанные на использовании нечеткой логики, позволяющей учитывать различные типы неопределенностей и кластеризовать пользовательские профили [1].

Самым распространенным подходом при реализации рекомендательных систем в электронной коммерции является метод коллаборативной фильтрации (collaborative filtering). Данный подход позволяет вырабатывать рекомендации, основанные на модели предшествующего поведения

Труды XVIII Международной конференции DAMDID/RCDL'2016 «Аналитика и управление данными в областях с интенсивным использованием данных», Ершово, 11-14 октября 2016

пользователей с учетом поведения других пользователей со сходными характеристиками [2]. Для выделения групп пользователей со сходными характеристиками, как правило, используются различные методы извлечения данных (data mining), которые, в свою очередь, используют алгоритмы кластеризации. В частности, в работе [3] указывается, что решение вопросов идентификации групп пользователей по своей природе опирается на использование методов кластеризации. Кластеризация данных может быть также использована для генерации профилей пользователей на основе информации о действиях каждого пользователя, а затем для формирования групп пользователей на основе их профилей.

Основной задачей, которую авторы данной работы ставили перед собой, является разработка комбинированного подхода к построению рекомендательных систем, обеспечивающего наиболее полное использование всех данных о посетителях интернет-магазинов с целью выработки рекомендаций наиболее адекватно отражающих их ожидания (пертинентность предложения). Научно практическая новизна работы заключается в идее комбинированного использования методов Item-Item CF и User-User CF, что позволяет минимизировать недостатки каждого из них и добиться более высокого качества работы рекомендательной системы в целом. Данная статья входит в серию статей по данной проблематике и посвящена решению задачи коллаборативной фильтрации User-User CF с использованием методов кластеризации для выявления групп пользователей со схожими предпочтениями. Предполагается, что при наличии качественных данных о пользовательской активности метод User-User CF дает наиболее адекватные прогнозы. Использование методов кластеризации для решения задачи выявления групп пользователей со схожими характеристиками позволяет добиться хорошего быстродействия и качества работы алгоритма.

1 Персонализация контента

Практически все современные интернет-ресурсы, ориентированные на работу с большим количеством пользователей, собирают информацию об их активности и анализируют (обрабатывают) ее с целью персонализации своего контента для каждого конкретного пользователя. В качестве наиболее характерных примеров можно выделить:

1. поисковые машины, которые собирают и систематизируют информацию о страницах в сети Интернет заинтересовавших конкретных пользователей;
2. интернет-магазины, которые собирают и систематизируют сведения о предпочтениях своих пользователей в части товаров и услуг;
3. форумы, интернет-дневники и социальные сети, которые собирают информацию о том, в каких тематических разделах и группах принимает участие каждый пользователь и насколько активно.

Другими достаточно распространенными примерами протоколирования действий пользователей являются счетчики посещений страниц, прокси-серверы и интернет-провайдеры.

Собранные данные о пользовательской активности характеризуются большим объемом и разнородностью. Традиционные базы данных малоприменимы для работы с этими данными по причине больших объемов данных и повышенных требований к производительности [4]. Как правило используются так называемые NoSQL системы управления данными (HBase, Cassandra). Их характерными особенностями являются отказ от транзакций, практически линейная масштабируемость, высокая скорость обработки запросов, отсутствие жесткой схемы данных.

В контексте проблемы персонализации контента (а также прогнозирования, выявления предпочтений и групп схожих ресурсов) встает задача обработки этих данных и выявления определенных закономерностей, позволяющих сделать выводы о конкретных предпочтениях пользователей. Таким образом, основной целью обработки данных о пользовательской активности является извлечение полезной информации, которая может, в свою очередь, использоваться для решения следующих задач [5]:

1. Кластеризация ресурсов. Группирование схожих по множеству посетителей ресурсов в несколько кластеров (групп) ресурсов. Кластеризация позволяет строить каталоги ресурсов, а также выявлять недостатки существующих тематических каталогов.
2. Кластеризация пользователей. Группирование схожих пользователей в кластеры аналогично кластеризации ресурсов. Позволяет выявлять группы пользователей со схожими интересами.
3. Построение устойчивых поведенческих профилей пользователей в виде перечня групп ресурсов, посещаемых как данным пользователем, так и схожими с ним пользователями.
4. Построение расширенных профилей пользователей, включающих социально-демографические данные (анкеты), описательные статистики и поведенческие профили. Расширенные профили позволяют классифицировать новых пользователей, выявлять зависимости между пользовательским поведением и социально-демографическими характеристиками.
5. Сегментация клиентской базы на основе расширенных профилей позволяет выделять сегменты, как по анкетным данным клиентов, так и по их поведению. Эта

информация используется при маркетинговых исследованиях.

6. Прямой маркетинг. Предоставление рекламы и маркетинговых предложений конкретному пользователю на основе его поведенческого профиля.
7. Персонализация контента. Представление каждому пользователю сайта наиболее интересной для него информации в наиболее удобном для него виде. Знание информационных предпочтений пользователя позволяет динамически перестраивать контент сайта.
8. Построение карт сходства ресурсов и пользователей. Позволяет отображать множества наиболее посещаемых ресурсов и наиболее активных пользователей в виде точечного графика. Схожим ресурсам (пользователям) соответствуют близкие точки на карте. Карту сходства можно использовать как графическое средство навигации.

Существуют различные методы и подходы, используемые на практике при решении перечисленных выше задач. Весь класс этих методов принято называть методами коллаборативной фильтрации.

2 Коллаборативная фильтрация

Результатом первого этапа обработки данных о пользовательской активности является построение матрицы активности, которая может нести различную информацию о действиях пользователя. Это может быть бинарная информация о посещении или не посещении заданного ресурса данным пользователем, частота (или число) пользований ресурса r пользователем u, стоимость или рейтинг, проставленный пользователем u для ресурса r и т.д. Для оценки степени схожести пользователей в плане их предпочтений и построения поведенческих профилей могут использоваться различные функции сходства (метрики). Наиболее популярными среди них являются: косинусная мера, коэффициент корреляции Пирсона, евклидово расстояние, коэффициент Танимото, Манхэттенское расстояние и другие [6].

Для решения задачи коллаборативной фильтрации используются три основных подхода: основанный на соседстве (memory based), основанный на модели (model based) и гибридный подход (hybrid). Первый подход исторически появился первым и характеризуется как достаточно простой в плане реализации, а также эффективный с точки зрения производительности. Но рекомендации, вырабатываемые с помощью данного метода, являются наименее точными. Подход, основанный на моделях при выработке рекомендаций, использует такие методы как метод байесовских сетей, кластеризации, латентной семантической модели, сингулярное разложение, вероятностный латентный семантический анализ, скрытое распределение Дирихле и марковской процесс принятия решений на основе моделей. Данный подход имеет целый ряд преимуществ и характеризуется более высоким качеством рекомендаций по сравнению с первым подходом. Гибридный подход сочетает преимущества подходов, основанных на соседстве и моделях. Данный подход является наиболее эффективным с точки зрения качества предсказаний, но при этом наиболее сложный в реализации и наиболее требовательный к производительности аппаратной платформы.

Основными проблемами, связанными с реализацией и практическим использованием алгоритмов коллаборативной фильтрации, являются разреженность данных, проблема холодного старта и масштабируемость. Разреженность данных изначально присуща исходным данным, которые используются для построения тематических профилей пользователей (покупатели просматривают и(или) оценивают только ограниченное число товаров и (или) услуг). Тем самым качество рекомендаций может быть очень низким, особенно на начальных этапах эксплуатации рекомендательной системы (когда еще не накоплено достаточное количество данных о пользовательской активности). Проблема разреженности данных напрямую связана с проблемой холодного старта, когда рекомендательная система должна вырабатывать рекомендации, имея минимальное количество данных о пользовательских предпочтениях. Проблема масштабируемости становится особенно острой для крупных интернет-магазинов, продающих тысячи товаров миллионам покупателей. При таком количестве товаров и покупателей сложность алгоритма резко возрастает, что усугубляется тем фактом, что рекомендательная система должна давать результат в считанные секунды. Дополнительно к перечисленному выше можно добавить проблему ограничения разнообразия предложений. Рекомендательные системы, использующие коллаборативную фильтрацию, склонны предлагать товары, уже пользующиеся популярностью, что препятствует продвижению новых товаров и услуг [7].

3 Кластеризация пользовательских профилей

Одним из способов решения задачи коллаборативной фильтрации, успешно используемых при реализации рекомендательных систем в современных системах электронной коммерции, является кластеризация. В настоящее время кластеризация – объединение в группы схожих объектов – является одной из фундаментальных задач в области анализа данных и Data Mining [8]. Существует большое количество методов кластеризации, которые условно можно разбить на

следующие основные группы: использующие вероятностный подход (K-means, EM-алгоритм), использующие методы искусственного интеллекта (нейронные сети, генетические алгоритмы), использующие теоретико-графовые модели, иерархические алгоритмы. В контексте решения задачи коллаборативной фильтрации наибольшее распространение получили алгоритмы, использующие вероятностный подход и иерархические алгоритмы.

В рамках проводимых исследований по повышению пертинентности информации в рекомендательных системах поддержки жизнеобеспечения авторы работы оценили возможность применения всех приведённых выше методов кластеризации, а также их комбинации. В результате проведенного имитационного моделирования было установлено, что наиболее подходящим с учётом существующих для работы ограничений по скорости и объемам обработки данных для выявления групп пользователей со схожими предпочтениями (User-User CF) был признан метод кластеризации K-средних (K-means). В целом данный статистический метод прост в реализации и является хорошо масштабируемым [9]. Вычислительная сложность алгоритма O(nkl), где n – число объектов, k - число кластеров, l - число итераций. Одним из недостатков данного метода является необходимость заранее задавать число кластеров для разбиения. Для решения этой задачи проводится предварительный анализ исходных данных, позволяющий посредством минимизации суммы внутрикластерных расстояний выявить оптимальное число кластеров для разбиения.

Следующим шагом стала проверка метода кластеризации на тестовом массиве данных, предоставленных интернет-магазином Thaisoap. Магазин ориентирован на продажу натуральной тайской косметики и кокосового масла. Каталог товаров магазина содержит более
1 500 наименований товаров, которые разбиты на 180 классов (44 корневых классов, 136 подклассов). Ежедневно магазин посещают в среднем около
1 500 посетителей и проводят на нем (в среднем) порядка 11 минут каждый (на каждого посетителя приходится в среднем 28 переходов по ссылкам).

Таким образом исходными данными для построения матрицы активности пользователей стали log-файлы (размер файла с записями за неделю превышает 100 МБ). При этом из всех предоставленных неявных данных для публикации было решено выбрать вычисления по наиболее универсальному агрегированному показателю – число обращений к конкретной категории товара. Кластеризация по данному показателю позволяет сформировать рекомендацию постоянному пользователю сразу в двух случаях: и когда все товары в категории, к которой тяготеет пользователь, просмотрены (User-User переход между категориями), и когда нет (внутрикатегориальная персональная фильтрация).

В результате обработки поступивших данных для каждого пользователя в матрице активности было определено число обращений к конкретной категории товаров. Таким образом, каждая строка матрицы активности пользователей представляет собой вектор оценок, соответствующих различным категориям товаров (тематический профиль пользователя). Профиль пользователя характеризует степень его интереса к каждой группе товаров.

На рисунке 1 представлен фрагмент матрицы активности пользователей, построенный в результате обработки log-файла интернет-магазина Thaisoap за недельный период времени. Число строк таблицы превысило пять тысяч (количество уникальных посетителей за данный период времени), число столбцов 180 (категории товаров).

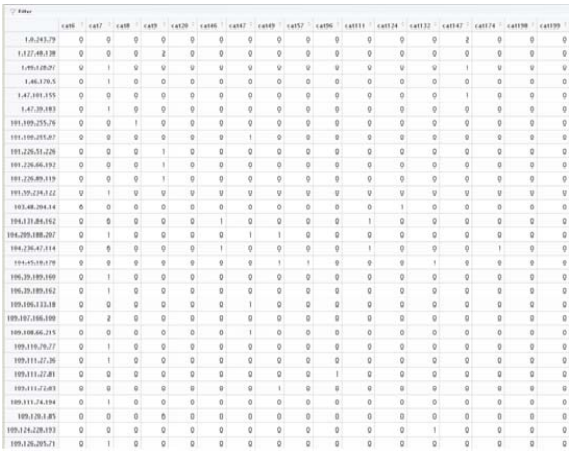


Рисунок 1 Матрица активности пользователей интернет-магазина Thaisoap

Предварительным этапом перед выявлением групп пользователей со схожими характеристиками методами кластеризации является вычисление попарных расстояний между элементами матрицы активности пользователей. В качестве метрики расстояния (функция сходства) было использовано расстояние Евклида (геометрическое расстояние в многомерном пространстве), которое является одной из наиболее простых для реализации и часто используемых на практике метрик на сегодняшний день.

На рисунке 2 представлена гистограмма расстояний, полученная в результате обработки матрицы с попарными расстояниями между объектами. Для построения гистограммы использовался статистический пакет R. Предварительное рассмотрение результатов позволяет сделать вывод, что предпочтения (вектора активности) пользователей интернет-магазина Thaisoap не сильно отличаются (т.е. скорее всего большинство пользователей интересуются схожими категориями товаров). Данный вывод можно сделать исходя из того, что около 90% данных на

гистограмме сосредоточено на отрезке расстояний от 0 до 5.

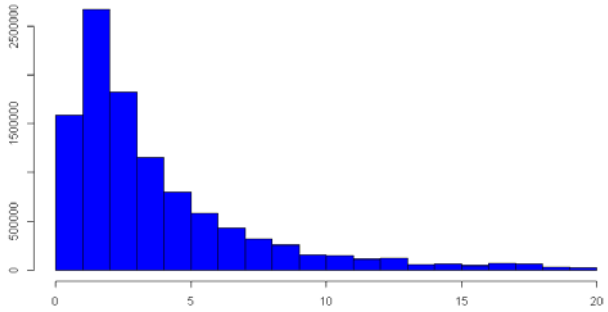


Рисунок 2 Гистограмма расстояний.

Следующим шагом было определение оптимального числа кластеров, требуемого для применения метода кластеризации К-средних. На рисунке 3 приведены результаты расчёта внутрикластерной суммы квадратов расстояний по методу локтя (Elbow method). Данный метод дал число в 30 кластеров.

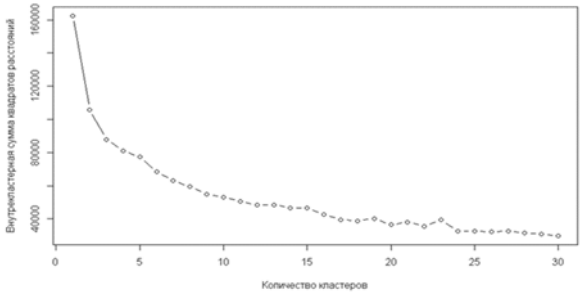


Рисунок 3 Анализ количества кластеров

Таким образом были получены все необходимые параметры и проведена кластеризация данных по матрице активности пользователей. На рисунке 4 в графическом виде представлен результат работы алгоритма кластеризации. Анализ полученных результатов позволяет выделить несколько наиболее крупных кластеров пользователей с номерами 1, 9, 20, 21 и 22. Пользователи из кластера под номером 1 (размер кластера 1 384) демонстрируют слабое предпочтение ко всем категориям товаров. Возможно, что в кластер с номером 1 попали посетители сайта, которые пришли без конкретной цели, например, просто ознакомиться с предлагаемым ассортиментом товаров. Пользователи из кластера номер 9 (размер 180) демонстрируют явное предпочтение к категории cat9 ("Лицо"). Пользователи из кластера номер 20 (размер 1 366) демонстрируют предпочтение к категории cat7 ("Кокосовое Масло"). Сумма квадратов расстояний относительно других кластеров мала, что говорит о небольшом различии объектов внутри кластера. Пользователи из кластера номер 21 (размер 622) демонстрирую также предпочтение к категории cat7 ("Кокосовое Масло"). Пользователи из кластера номер 22 (размер 180) демонстрирую предпочтение к категориям cat47 ("MUST HAVE Зима 2016") и cat49 ("ХИТЫ нашего магазина"). Проведение расчётов на основании данных за другие недели дало схожие результаты.

Таким образом использование метода К-средних на реальных данных с одной стороны подтвердило результаты имитационного моделирования, а с другой стороны выявило недостатки анализа только предпочтений групп пользователей со схожими интересами только по одному агрегированному неявному показателю (User-User CF): для большинства посетителей отсутствует возможность сформировать сколь-либо персональное информационное предложение и требуется как расширенный анализ оставшихся неявных данных, так и иных методов, например, коллаборативной фильтрации посредством анализа взаимосвязей между объектами (Item-Item CF).

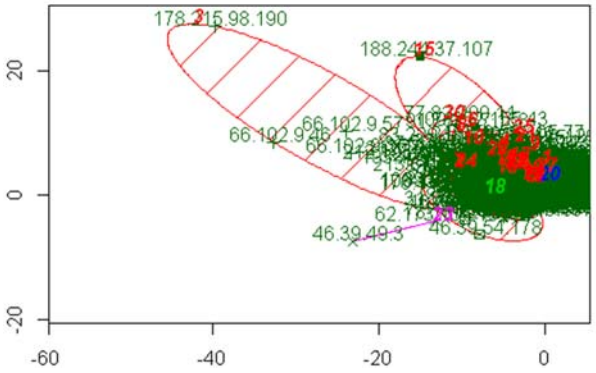


Рисунок 4 Результат работы алгоритма кластеризации

Заключение

В заключение необходимо отметить, что использование различных алгоритмов кластеризации для решения задачи коллаборативной фильтрации в настоящее время является одним из перспективных направлений. Большие объемы данных о пользовательской активности и высокие требования к быстродействию рекомендательных систем накладывают определенные ограничения на используемые алгоритмы. Поэтому наибольшее применение в этой области получили алгоритмы, основанные на оптимизации некоторой целевой функции, определяющей оптимальное (в контексте задачи) разбиение множества объектов на кластеры. В частности, большой популярностью пользуются алгоритмы семейства К-средних (K-means, fuzzy C-means, Густафсон-Кесселя), которые в качестве целевой функции используют сумму квадратов взвешенных отклонений координат объектов от центров искомых кластеров.

В статье на примере данных интернет-магазина Thaisoap показаны достоинства и недостатки кластеризации по простому агрегированному неявному показателю – числу обращений к конкретной категории товара с использованием алгоритма кластеризации К-средних (K-means). В том числе была подтверждена слабая применимость

метода в условиях холодного старта (для новых или малоактивных пользователей). В данном случае требуется применение иных неявных показателей или принципиально других методов, например, Item-Item CF, которые соответственно дают больше информации о пользователе за единицу времени или лучше работают в ситуациях, когда данные о пользовательской активности минимальны. При этом по мере накопления данных о предпочтениях пользователей при выработке рекомендаций рассмотренный метод начинает давать всё более уместные предложения и рекомендуется к использованию как основной.

Литература

[1] А.Н.Алфимцев, В.В. Девятков, С.А.Сакулин Персонализация в гипертекстовых сетях на основе распознавания действий пользователей и нечеткого агрегирования // Вестник МГТУ им.Баумана, Сер. «Приборостроение», 2012, №3.

[2] М. Тим Джонс Рекомендательные системы: Часть 1. Введение в подходы и алгоритмы // Статья в сети Интернет, URL: http://www.ibm.com/developerworks/ru/library/os-recommender1/, 2013.

[3] Марманис Х., Бабенко Д. Алгоритмы интеллектуального Интернета // СПб.-М.: Символ, 2011. – 466 с.

[4] С.А.Филиппов, В.Н.Захаров, С.А.Ступников, Д.Ю.Ковалев Организация больших объемов данных в рекомендательных системах поддержки жизнеобеспечения, входящих в состав глобальных платформ электронной коммерции // XVII международная конференция «Аналитика и управление данными в областях с интенсивным использованием данных» DAMDID/RCDL'2015. Обнинск, 2015.

[5] В.А. Лексин Технология персонализации на основе выявления тематических профилей пользователей и ресурсов Интернет // ВКР Магистра, Вычислительный Центр им. А.А. Дородницина РАН, 2007.

[6] Xiaoyuan Su, Taghi M. Khoshgoftaar A survey of collaborative filtering techniques // Advances in Artificial Intelligence, Volume 2009 (2009), Article ID 421425, 19p.

[7] Fleder D., Hosanagar K. Blockbuster Culture's Next Rise or Fall: The Impact of Recommender Systems on Sales Diversity // Management Science, Vol. 55, No. 5, May 2009, pp. 697-712.

[8] Барсегян и др. Методы и модели анализа данных: OLAP и Data Mining. – СПб., 2004.

[9] Adam Coates and Andrew Y. Ng. Learning Feature Representations with K-means // Stanford University, 2012, Статья в сети Интернет, URL: http://www.cs.stanford.edu/~acoates/papers/coatesng_nntot2012.pdf.

Clustering of user profiles based on real implicit data in e-commerce recommender systems

Stanislav A. Philippov, Victor N. Zaharov, Sergey A. Stupnikov, Dmitriy Yu. Kovalev

This work is devoted to description of key tasks in the context of building the online store information systems. The main objective is to identify user preferences and their formalization through the formation of the users' behavioral profiles followed by the identification of user groups with similar characteristics. The main source of information about user preferences is implicitly an array of data collected about their actions when navigating through the pages of online shopping. The authors present an approach to building a recommendation system based on collaborative filtering problem solving using cluster analysis techniques to identify groups of users with similar preferences. Advantages of the solution are demonstrated on the example of test data set obtained from the current online store Thaisoap. A unique identifier of the project supported by the Ministry of education and science of the RF is RFMEFI60414X0139.

Метод определения подобия информационных единиц по неявным пользовательским предпочтениям в рекомендательных системах поддержки жизнеобеспечения

© С. А. Филиппов © В. Н. Захаров © С. А. Ступников
© Д. Ю. Ковалев
Институт проблем информатики ФИЦ ИУ РАН,
Москва
stanislav@philippov.ru VZakharov@ipiran.ru ssa@ipi.ac.ru
dm.kovalev@gmail.com

Аннотация

Целью данной работы является описание метода определения подобия информационных единиц посредством анализа данных о пользовательских предпочтениях. Метод является реализацией подхода Item-Item CF (коллаборативная фильтрация на основе подобия информационных единиц), который в свою очередь является одним из наиболее популярных подходов к построению современных рекомендательных систем. Исходными данными для коллаборативной фильтрации (другими словами для выявления пользовательских предпочтений) являются данные о пользовательской активности при просмотре страниц конкретных интернет-ресурсов (информационных единиц). Данные могут собираться как явным (оценки, опросы, рейтинги), так и неявным образом (протоколирование действий пользователей). Предложенный метод позволяет решить проблему холодного старта, т.е. выдачи рекомендаций в период отсутствия подробной информации о посетителе системы поддержки жизнеобеспечения (здесь и далее под такой системой подразумевается интернет-магазин), но при наличии неявных данных о маршрутах других посетителей системы. Метод опробован на реальных данных, полученных с действующего интернет-магазина Thaisoap, где подтвердил возможность своей применимости в рамках поставленной задачи. Работа выполнена при поддержке Министерства образования и науки РФ, уникальный идентификатор проекта RFMEFI60414X0139.

Труды XVIII Международной конференции DAMDID/RCDL'2016 «Аналитика и управление данными в областях с интенсивным использованием данных», Ершово, 11-14 октября 2016

Введение

Одним из современных трендов в развитии Интернет является персонализация. Поисковые системы, социальные сети, форумы, новостные ресурсы и Интернет магазины стараются адаптировать внешний вид и содержимое (контент) своих страниц под нужды конкретных пользователей. По результатам исследования компании Evergage (www.evergage.com) в 2015 году персонализацию в реальном времени использовали 44% веб сайтов, 17% мобильных сайтов, 13% веб-приложений и 9% мобильных приложений [1]. При этом 78% тех, кто не использует персонализацию сейчас, утверждают, что планируют начать в течение следующих 12 месяцев. Увеличение вовлеченности посетителей, улучшение пользовательского опыта и повышение конверсии считаются самыми важными результатами ее применения.

Предоставление персонализированного контента пользователям позволяет существенно повысить эффективность сайтов, которая выражается в терминологии маркетинга таким показателем как конверсия (число посетителей, совершивших полезные действия к общему числу посетителей выраженное в процентах). Для качественной персонализации сайтов, ориентированных на работу с большой аудиторией пользователей, как правило, используется комплексный подход, сочетающий маркетинговые исследования и анализ поведения конкретных посетителей сайтов. Информацию о маркетинговых качествах посетителей можно получить, в том числе используя системы веб-аналитики, такие как Adobe Digital Marketing Suite или Google Analytics и Siteapps.com. Исходными данными для анализа поведения пользователей являются сведения об их активности, которые могут собираться явным или неявным образом. Явным образом получают результаты голосований и опросов, а также оценки, которые пользователи дают

тем или иным объектам на сайтах. Основное количество информации о пользовательской активности собирается неявным образом посредством протоколирования его действий. Предметом отслеживания являются переходы пользователей по ссылкам на сайтах, время их пребывания на отдельных страницах, факты покупки товаров и услуг. Необходимо отметить, что, речь идет об огромных массивах данных, которые являются неоднородными и требующими отдельных подходов к интерпретации.

В сфере электронной коммерции основным инструментом персонализации контента являются рекомендательные системы, обеспечивающие автоматическую обработку данных о пользовательской активности и выработку рекомендаций на товары и услуги, которые могут быть интересны конкретным пользователям. При реализации рекомендательных систем широко используются методы интеллектуального анализа данных (Data Mining) [2].

Основной задачей, которую авторы данной работы ставили перед собой, является разработка комбинированного подхода к построению рекомендательных систем, обеспечивающего наиболее полное использование всех данных о посетителях интернет-магазинов с целью выработки рекомендаций, наиболее адекватно отражающих их ожидания (пертинентность предложения). Научно практическая новизна работы заключается в идее комбинированного использования методов Item-Item CF и User-User CF, что позволяет минимизировать недостатки каждого из них и добиться более высокого качества работы рекомендательной системы в целом. Данная статья входит в серию статей по данной проблематике и посвящена описанию метода определения подобия информационных единиц по неявным пользовательским предпочтениям, который является вариантом реализации метода Item-Item CF. Данный метод позволяет вырабатывать приемлемые по качеству рекомендации в условиях, когда сведения о пользовательских предпочтениях отсутствуют, минимальны или слабо информативны. Для выявления групп подобных товаров используются методы кластеризации, что позволяет добиться хороших показателей качества и быстродействия в работе алгоритма.

1 Построение рекомендательных систем с использованием методов коллаборативной фильтрации

Основная задача рекомендательной интернет-системы – формирование контента, максимально соответствующего ожиданиям, в том числе неявным, конкретного пользователя. Для решения этой задачи в большинстве современных рекомендательных систем используется один из двух базовых подходов: коллаборативная фильтрация (collaborative filtering, CF) и контентная фильтрация (content-based filtering, CbF) [3].

Метод контентной фильтрации фокусируется на выявлении объектов со схожими характеристиками по отношению к тем объектам, которые уже заинтересовали пользователя. При этом учитывается модель поведения пользователя и характеристики (контент) заинтересовавших его объектов. При выработке рекомендаций выявляются объекты со схожими характеристиками (контентом). Для эффективной работы метода контентной фильтрации, как правило, необходимо подробное описание характеристик объектов (так в проекте Music Genome Project музыкальный аналитик оценивает каждую композицию по сотням различных музыкальных характеристик), а также сведения о конкретном пользователе (например, ответы на конкретные вопросы в анкете).

В основе метода коллаборативной фильтрации лежит предположении о консервативности пользовательских предпочтений (т.е. пользователи, одинаково оценивающие определенные объекты, скорее всего аналогичным образом будут оценивать и новые объекты со сходными характеристиками) [4]. По существу, рекомендации базируются на автоматическом сотрудничестве множества пользователей и на выделении (методом фильтрации) тех пользователей, которые демонстрируют схожие предпочтения или шаблоны поведения. Таким образом, метод коллаборативной фильтрации вырабатывает рекомендации, основанные на модели предшествующего поведения пользователя и с учетом поведения пользователей со схожими характеристиками.

Наибольшее распространение в сфере электронной коммерции получили рекомендательные системы, использующие следующие варианты реализации метода коллаборативной фильтрации, а также их гибриды:

- коллаборативная фильтрация посредством анализа предпочтений групп пользователей со схожими интересами (User-User Collaborative Filtering, User-User CF);
- коллаборативная фильтрация посредством анализа взаимосвязей между объектами (Item-Item Collaborative Filtering, Item-Item CF);

Основными проблемами, связанными с реализацией и практическим использованием алгоритмов коллаборативной фильтрации, являются разреженность данных, проблема холодного старта и масштабируемость. Дополнительно к перечисленным проблемам можно отметить проблему ограничения разнообразия предложений. Рекомендательные системы, использующие коллаборативную фильтрацию, склонны предлагать товары уже пользующиеся популярностью, что создает проблемы для продвижения новых товаров и услуг [5].

В методе User-User CF определяется сходство между пользователями и в качестве рекомендаций пользователю выдается n самых часто покупаемых товаров k наиболее похожими на него покупателями. Для оценки степени схожести пользователей в плане их предпочтений могут использоваться различные функции сходства (метрики). Наиболее популярными среди них являются: евклидово расстояние, косинусная мера, расстояние Хэмминга, коэффициент корреляции Пирсона, коэффициент Танимото, Манхэттенское расстояние и некоторые другие [4, 6]. Определение рекомендаций методом User-User CF предполагает построение матрицы активности пользователей, каждая строка которой описывает действия конкретного пользователя применительно к конкретному объекту (категория, товар, услуга) на сайте. Действия пользователей могут обозначаться самыми различными способами. Например, это может быть бинарная информация о посещении или не посещении заданного ресурса данным пользователем, частота (или число) пользований ресурса r пользователем u, стоимость или рейтинг, проставленный пользователем u для ресурса r и т.д. Таким образом, каждая строка матрицы активности представляет собой вектор оценок, соответствующих различным категориям товаров (тематический профиль пользователя). Профиль пользователя характеризует степень его интереса к каждой группе товаров. Для каждой пары «пользователь-объект (товар, услуга, действие)» в матрице активности вычисляется мера близости с использованием выбранной метрики [7].

Для поиска рекомендаций конкретному пользователю на основании его поведенческого профиля используются три основных подхода: основанный на соседстве (memory based), основанный на модели (model based) и гибридный подход (hybrid). В современных коммерческих системах наибольшее распространение получили гибридный подход и подход, основанный на использовании моделей (алгоритмы кластеризации, байесовские сети доверия, латентные семантические модели) [3, 9]. Для выявления групп пользователей со схожими характеристиками часто используются различные алгоритмы кластеризации.

Метод Item-Item CF исторически появился как альтернатива методу User-User CF, призванная повысить производительность рекомендательных систем для тех магазинов, где число покупателей существенно превышает количество наименований товаров в каталоге [8]. Первоначально данный метод был предложен компанией Amazon для решения следующих основных проблем подхода User-User CF: проблема холодного старта и проблема частого обновления данных о пользовательской активности. Проблема холодного старта существенно снижает качество работы рекомендательной системы вследствие отсутствия данных о предпочтениях новых (или мало активных) пользователей. Проблема частого обновления данных о пользовательской активности (в случае компании Amazon речь идет о миллионах покупателей) резко снижает производительность рекомендательной системы в целом.

Основная идея метода Item-Item CF заключается в группировке информационных единиц (товары, услуги, действия) имеющих сходные оценки пользователей (рейтинги). Рекомендации вырабатываются по следующему принципу: пользователю оценившему объект X высоко будет предложен объект Y, который высоко оценили другие пользователи, также высоко оценившие и объект X. Использование метода Item-Item CF позволяет повысить качество рекомендаций для новых пользователей (нет критической зависимости от данных о пользовательских предпочтениях), а также значительно повышает производительность рекомендательной системы в случае, когда количество пользователей существенно превышает количество объектов (характеристики объектов меняются реже). При этом качество рекомендаций в среднем выше, чем в случае использования подхода, основанного на анализе пользовательских профилей. Для вычисления попарной близости информационных единиц могут использоваться те же метрики, что и в случае с парами «пользователь-объект» (часто используется косинусная или модифицированная косинусная меры). Для поиска рекомендаций на основании матрицы объектов часто используются весовые функции и методы регрессионного анализа. Одним из перспективных методов решения задачи Item-Item CF является метод Item2Vec [10]. Тем не менее для большинства интернет-магазинов подход, связанный с рекомендациями по рейтингам, слабо применим в силу отсутствия возможности мотивировать пользователей определять рейтинг информационных единиц (покупатели приходят из поисковых систем и товарных каталогов, делают нужную им покупку и уходят, чтобы больше никогда не вернуться). И встает задача, как в таких условиях сделать рекомендацию (информационное предложение), на которую откликнется пользователь.

2 Определение подобия (кластеров) информационных единиц по неявным пользовательским предпочтениям

В целях решения задачи формирования рекомендации с уместной информацией в условиях недостаточности знаний о пристрастиях пользователей авторами предлагается использовать метод, в основе которого лежит расчёт близости пар и последующая группировка (кластеризация) информационных единиц на основе данных пользователей, последовательно просматривающих несколько товаров. При отсутствии данных предлагается использовать обычные классификаторы с учётом цены и параметров объектов, список «Новинки», а также матрицу «С этим товаром покупают» (аксессуары,

дополняющие основную покупку). При данном подходе явное участие пользователей интернет-магазина в формировании рейтинга товаров не требуется.

Первым шагом алгоритма является построение матрицы подобия информационных единиц, где и по вертикали, и по горизонтали присутствуют все информационные единицы интернет-магазина. Заполнение матрицы происходит по следующему правилу: если пользователь последовательно просмотрел два товара, то вес подобия в матрице для этих двух товаров увеличивается на 1.

Для обработки матрицы в целях выявления групп информационных единиц, которые являются близкими по своим оценкам подобия, из всех известных алгоритмов кластеризации в результате проведённого моделирования был выбран современный производительный алгоритм Affinity Propagation. Одним из преимуществ данного алгоритма является отсутствие необходимости предварительной оценки оптимального количества кластеров [11].

Приведённый метод кластеризации был опробован на тестовом массиве данных, предоставленных интернет-магазином Thaisoap. Магазин ориентирован на продажу натуральной тайской косметики и кокосового масла. Каталог товаров магазина содержит более 1 500 наименований товаров, которые разбиты на 180 классов (44 корневых классов, 136 подклассов). Ежедневно магазин посещают в среднем около 1 500 посетителей и проводят на нем (в среднем) порядка 11 минут каждый (на каждого посетителя приходится в среднем 28 переходов по ссылкам). Исходные данные охватывают период в один квартал (IV квартал 2015 года), в котором каталог товаров был неизменен.

	p43	p44	p45	p46	p47	p48	p49	p50	p51	p52	p53	p54
p43	1.0000000	0.04347826	0.00000000	0.04347826	0.00000000	0.04347826	0.04347826	0.04347826	0.30434783	0.04347826	0.08695652	0.00000000
p44	0.04347826	1.00000000	0.00000000	0.04347826	0.00000000	0.00000000	0.08695652	0.04347826	0.04347826	0.08695652	0.13043478	0.04347826
p45	0.00000000	0.00000000	1.00000000	0.00000000	0.00000000	0.00000000	0.00000000	0.00000000	0.00000000	0.00000000	0.00000000	0.00000000
p46	0.04347826	0.04347826	0.00000000	1.00000000	0.00000000	0.00000000	0.04347826	0.04347826	0.04347826	0.04347826	0.04347826	0.00000000
p47	0.00000000	0.00000000	0.00000000	0.00000000	1.00000000	0.00000000	0.00000000	0.00000000	0.00000000	0.00000000	0.00000000	0.00000000
p48	0.04347826	0.00000000	0.00000000	0.00000000	0.00000000	1.00000000	0.13043478	0.00000000	0.00000000	0.00000000	0.08695652	0.00000000
p49	0.04347826	0.08695652	0.00000000	0.04347826	0.00000000	0.13043478	1.00000000	0.04347826	0.08695652	0.04347826	0.08695652	0.00000000
p50	0.04347826	0.04347826	0.00000000	0.04347826	0.00000000	0.00000000	0.04347826	1.00000000	0.04347826	0.04347826	0.04347826	0.00000000
p51	0.30434783	0.04347826	0.00000000	0.04347826	0.00000000	0.00000000	0.08695652	0.04347826	1.00000000	0.17391304	0.13043478	0.00000000
p52	0.04347826	0.08695652	0.00000000	0.04347826	0.00000000	0.00000000	0.04347826	0.04347826	0.17391304	1.00000000	0.30434783	0.00000000
p53	0.08695652	0.13043478	0.00000000	0.04347826	0.00000000	0.08695652	0.08695652	0.04347826	0.13043478	0.30434783	1.00000000	0.00000000
p54	0.00000000	0.04347826	0.00000000	0.00000000	0.00000000	0.00000000	0.00000000	0.00000000	0.00000000	0.00000000	0.00000000	1.00000000
p55	0.00000000	0.00000000	0.00000000	0.04347826	0.04347826	0.00000000	0.00000000	0.00000000	0.00000000	0.00000000	0.00000000	0.00000000
p56	0.00000000	0.00000000	0.00000000	0.00000000	0.00000000	0.00000000	0.00000000	0.00000000	0.00000000	0.00000000	0.00000000	0.00000000
p58	0.17391304	0.00000000	0.00000000	0.00000000	0.00000000	0.00000000	0.08695652	0.00000000	0.00000000	0.00000000	0.08695652	0.00000000
p60	0.04347826	0.08695652	0.04347826	0.08695652	0.00000000	0.00000000	0.04347826	0.04347826	0.04347826	0.04347826	0.08695652	0.00000000
p61	0.00000000	0.00000000	0.00000000	0.00000000	0.00000000	0.00000000	0.00000000	0.00000000	0.00000000	0.00000000	0.00000000	0.00000000
p64	0.00000000	0.00000000	0.00000000	0.00000000	0.00000000	0.00000000	0.00000000	0.00000000	0.00000000	0.00000000	0.00000000	0.00000000
p65	0.08695652	0.04347826	0.00000000	0.04347826	0.00000000	0.04347826	0.04347826	0.04347826	0.08695652	0.08695652	0.08695652	0.00000000
p66	0.04347826	0.04347826	0.00000000	0.04347826	0.00000000	0.00000000	0.08695652	0.04347826	0.04347826	0.08695652	0.04347826	0.00000000
p67	0.04347826	0.04347826	0.00000000	0.04347826	0.00000000	0.00000000	0.04347826	0.04347826	0.08695652	0.04347826	0.04347826	0.00000000
p71	0.00000000	0.00000000	0.00000000	0.00000000	0.00000000	0.00000000	0.00000000	0.00000000	0.00000000	0.00000000	0.00000000	0.00000000
p72	0.00000000	0.00000000	0.00000000	0.00000000	0.00000000	0.00000000	0.00000000	0.00000000	0.00000000	0.00000000	0.00000000	0.00000000
p73	0.00000000	0.00000000	0.00000000	0.00000000	0.00000000	0.00000000	0.00000000	0.00000000	0.00000000	0.00000000	0.00000000	0.00000000
p75	0.08695652	0.04347826	0.00000000	0.04347826	0.00000000	0.00000000	0.04347826	0.04347826	0.13043478	0.08695652	0.08695652	0.00000000

Рисунок 1 Матрица подобия товаров

На основе указанных данных была построена матрица подобия по всему временному периоду. На рисунке 1 представлен фрагмент получившейся матрицы подобия товаров для всех товаров из каталога (значения нормированы). Всего в каталоге на данный момент присутствует 1522 товара. Как видно из рисунка матрица сильно разряжена, так как для многих пар товаров оценка подобия отсутствует (т.е. в течение анализируемого периода времени пользователи не интересовались некоторыми товарами из каталога).

В результате обработки матрицы подобия по алгоритму Affinity Propagation (с использованием статистического пакета R) была построена гистограмма расстояний. Результаты работы алгоритма представлены на рисунке 2 в виде кластерной тепловой карты (размерность карты 1522 на 1522). Преобладание одного цвета на карте обусловлено тем фактом, что в тестовой выборке данных для большинства пар товаров не определена оценка подобия (т.е. пользователи не интересовались данными товарами в течение рассматриваемого в тестовой выборке периода времени).

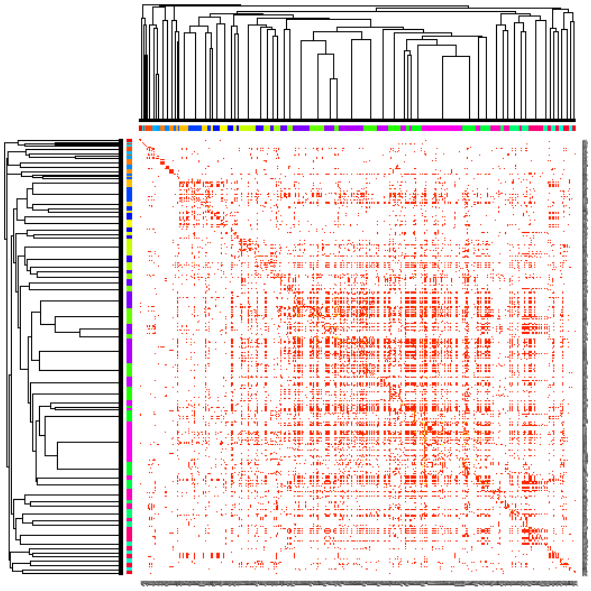


Рисунок 2 Кластерная тепловая карта

Всего алгоритм выделил 64 кластера, наиболее крупными из которых являются кластера с номерами 5 (75 объектов), 8 (44 объекта),
10 (30 объектов), 19 (27 объектов) и 55 (31 объект).

Качество работы алгоритма можно оценить на примере кластера номер 5, описание которого представлено в таблице 1. В частности, видно, что для референсной информационной единицы (массажное кокосовое масло) в кластер подобия попали товары на основе кокосового масла или косвенно ассоциирующиеся с кремами и маслами для ухода за телом.

Аналогичные результаты показывает и исследование других полученных кластеров. Таким образом, метод определение подобия информационных единиц выполняет возложенные на него задачи: формируется рекомендация из информационных единиц (товаров), уместных по отношению к товару, который заинтересовал неизвестного посетителя в данный конкретный момент времени.

Таблица 1 Детализация кластера номер 5

Кластер	Референсная информационная единица	Примеры товаров из кластера
ID: 5 Size: 75	ID: 76. Нерафинированное 100% массажное кокосовое масло "Citronella" Tropicana, 100 мл.	ID: 43. Кокосовое масло Tropicana 1 литр, нерафинированное ID: 51. Кокосовое масло нерафинированное Tropicana в аптекарском флаконе, 90 мл. ID: 466. Восстанавливающий кокосовый ЛОСЬОН для тела Tropicana "Sweet Coconut" (без парабенов), 200 мл. ID: 624. Маска-эксфолиант для лица "Морской коллаген" Artiscent, 100 мл. ID: 1234. Мини-набор Шампунь и Кондиционер для волос "Золотой шелк с экстрактом шелковицы"

Заключение

Персонализация контента интернет-ресурсов на сегодня является одним из активно развивающихся направлений ИТ-индустрии. Важнейшими результатами ее применения являются увеличение вовлеченности посетителей, улучшение пользовательского опыта и повышение конверсии. Персонализация контента в сфере электронной коммерции выражается в адресном предложении товаров, а также услуг конкретным пользователям и реализуется посредством рекомендательных систем. Современные рекомендательные системы обеспечивают обработку огромных массивов данных о пользовательской активности с целью формирования предсказаний для конкретных пользователей в момент запроса.

В данной работе изложен метод определения подобия информационных единиц по неявным пользовательским предпочтениям в рекомендательных системах поддержки жизнеобеспечения на основе упрощенной метрики близости пар информационных единиц по алгоритму кластеризации Affinity Propagation. Метод проверен на данных интернет-магазина Thaisoap и показал по результатам высокий уровень уместности информации в формируемой рекомендации.

Таким образом описанный метод класса Item-Item CF вполне применим для новых (или малоактивных) пользователей. При этом по мере накопления данных о предпочтениях пользователей рекомендуются отдавать большее предпочтение методам класса User-User CF, которые дают тем более точные предсказания чем более подробны данные о пользовательской активности.

Литература

[1] Почему персонализация контента это еще не веб-персонализация // Статья в сети Интернет, URL: http://lpgenerator.ru/blog/2016/03/19/ pochemu-personalizaciya-kontenta-eto-eshe-ne-veb-personalizaciya/

[2] С.А.Филиппов, В.Н.Захаров, С.А.Ступников, Д.Ю.Ковалев Подходы к повышению пертинентности информационного предложения в медиасервисах на основе обработки больших объемов данных // XVII международная конференция «Аналитика и управление данными в областях с интенсивным использованием данных» DAMDID/RCDL’2015, Октябрь 13-16, Обнинск, 2015, с. 224-228..

[3] М. Тим Джонс Рекомендательные системы: Часть 1. Введение в подходы и алгоритмы // Статья в сети Интернет, URL: http://www.ibm.com/developerworks/ru/library/os-recommender1/, 2013.

[4] Xiaoyuan Su, Taghi M. Khoshgoftaar A survey of collaborative filtering techniques // Advances in Artificial Intelligence, Volume 2009 (2009), Article ID 421425, 19p.

[5] Fleder D., Hosanagar K. Blockbuster Culture's Next Rise or Fall: The Impact of Recommender Systems on Sales Diversity // Management Science, Vol. 55, No. 5, May 2009, pp. 697-712.

[6] В.А. Лексин Технология персонализации на основе выявления тематических профилей пользователей и ресурсов Интернет // ВКР Магистра, Вычислительный Центр им. А.А. Дородницина РАН, 2007.

[7] Брейкин Е. А. Рекомендательная система на основе коллаборативной фильтрации // Молодой ученый. — 2015. — №13. — С. 31-33.

[8] Greg Linden, Brent Smith and Jeremy York Amazon.com recommendations: Item-to-Item Collaborative Filtering // Industry Report, IEEE INTERNET COMPUTING, 2003.

[9] Барсегян А.А., Куприянов М.С., Степаненко В.В., Холод И.И. Методы и модели анализа данных: OLAP и Data Mining // СПб.: БХВ-Петербург, 2004. — 336 с.

[10] Barkan O., Koenigstein N. Item2Vec: Neural Item Embedding for Collaborative Filtering // arXiv preprint arXiv:1603.04259, Mar 2016.

[11] Brendan J. Frey, Delbert Dueck Clustering by passing messages between data points // Science 16 Feb 2007 Vol. 315, Issue 5814, pp. 972-976, DOI: 10.1126/science.1136800

Determination of similarity of information items based on implicit user preferences in life-support recommender systems

Stanislav A. Philippov, Victor N. Zaharov,
Sergey A. Stupnikov, Dmitriy Yu. Kovalev

The purpose of this paper is to describe the method for determining the similarity of the information items through the analysis of user preference data. The method is an implementation approach known as Item-Item CF (collaborative filtering based on the similarity of the information items), which in turn is one of the most popular approaches to the construction of modern recommender systems. Initial data for collaborative filtering are the data about users' activity when they are browsing web resources. Data can be collected as explicit (evaluations, surveys, ratings) and implicit (logging of users' actions). The proposed method solves the problem of cold start using implicit data about the routes of other users. The method was tested on real data from existing online store Thaisoap, which confirmed the possibility of its applicability in the framework of the task. A unique identifier of the project supported by the Ministry of education and science of the RF is RFMEFI60414X0139.

Исследовательские инфраструктуры в астрофизике

Research Data Infrastructures in Astrophysics

Accessing distributed computing resources by scientific communities using DIRAC services

V.Korenkov I.Pelevanyuk P.Zrelov
Joint Institute for Nuclear Research, Dubna
Plekhanov Russian Economics University, Moscow
korenkov@jinr.ru pelevanyuk@jinr.ru zrelov@jinr.ru

A.Tsaregorodtsev
CPPM, Aix Marseille University, CNRS/IN2P3, Marseille
Plekhanov Russian Economics University, Moscow
atsareg@in2p3.fr

Abstract

Scientific data intensive applications requiring simultaneous use of large amounts of computing resources are becoming quite common. This domain was pioneered by High Energy Physics (HEP) experiments at the LHC collider at CERN. However, researchers in other branches of science start to have similar requirements. The experience and software tools accumulated in the HEP experiments can be very valuable for these scientific communities. One of the software toolkits developed for building distributed computing systems is the DIRAC interware. It allows seamless integration into a single coherent system of computing and storage resources based on different technologies. This product was very successful to solve problems of large HEP experiments and was reworked in order to offer a general-purpose solution suitable for other scientific domains. Services based on the DIRAC interware are now proposed to users of several distributed computing infrastructures on the national and European levels. This significantly lowers the threshold to start working with large scale distributed computing systems for the new researchers.

1 Introduction

Large High Energy Physics experiments, especially those running at the LHC collider at CERN, have pioneered the era of very data intensive applications. The aggregated data volume of these experiments exceeds by today 100 PetaBytes, which includes both data acquired from the experimental setup as well as results of the detailed modeling of the detectors. Production and processing of these data required creation of a special distributed computing infrastructure - Worldwide LHC Computing Grid (WLCG). This is the first example of a large-scale grid system successfully used for a large scientific community. It includes more than 150 sites from more than 40 countries around the world. The sites altogether are providing unprecedented computing power and storage volumes. WLCG played a very important role in the success of the LHC experiments that achieved many spectacular scientific results like discovery of the Higgs boson, discovery of the pentaquark particle states, discovery of rare decays of B-mesons, and many others.

In order to create and operate the WLCG infrastructure, special software, so-called middleware, was developed to give uniform access to various sites providing computational and storage resources for the LHC experiments. Multiple services were deployed at the sites and centrally at CERN to ensure coherent work of the infrastructure, with comprehensive monitoring and accounting tools. All the communications between various services and clients are following strict security rules; users are grouped into virtual organizations with clear access rights to different services and with clear policies of usage of the common resources.

On top of the standard middleware that allowed building the common WLCG infrastructure, each LHC experiment, ATLAS, CMS, ALICE and LHCb, developed there own systems in order to manage their workflows and data and cover use cases not addressed by the middleware. Those systems have many similar solutions and design choices but are all developed independently, in different development environments and have different software architectures. This software is used to cope with large numbers of computational tasks and with large number of distributed file replicas by automation of recurrent tasks, automated data validation and recovery procedures. With time, the LHC experiments gained access also to other computing resources than WLCG. An important functionality provided by the experiments software layer is access to heterogeneous computing and storage resources provided by other grid systems, cloud systems and standalone large computing centers, which are not incorporated in any distributed computing network. Therefore, this kind of software is often called interware as it interconnects users and various computing resources

Proceedings of the XVIII International Conference «Data Analytics and Management in Data Intensive Domains» (DAMDID/RCDL'2016), Ershovo, Russia, October 11 - 14, 2016

and allows for interoperability of otherwise heterogeneous computing clusters.

Nowadays, other scientific domains are quickly developing data intensive applications requiring enormous computing power. The experience and software tools accumulated by the LHC experiments can be very useful for these communities and can save a lot of time and effort. One of the experiment interware systems, the DIRAC project of the LHCb experiment, was reorganized to provide a general-purpose toolkit to build distributed computing systems for scientific applications with high data requirements [1]. All the experiment specific parts were separated into a number of extensions, while the core software libraries are providing components for the most common tasks: intensive workload and data management using distributed heterogeneous computing and storage resources. This allowed offering the DIRAC software to other user communities and now it is used in multiple large experiments in high energy physics, astrophysics and other domains. However, for relatively small user groups with little expertise in distributed computing, running dedicated DIRAC services is a very difficult task. Therefore, several computing infrastructure projects are offering DIRAC services as part of their services portfolio. In particular, these services are provided by the European Grid Infrastructure (EGI) project. This allowed many relatively small user communities to have an easy access to a vast amount of resources, which they would never have otherwise.

Similar systems originating from other LHC experiments, like BigPanDa [14] or AliEn [15] were also offered to use by other scientific collaborations. However, their usage is more limited than the one of DIRAC. BigPanDa is providing mostly the workload management functionality for the users and is not supporting data management operations, whereas DIRAC is a complete solution for both types of tasks. AliEn provides support for both data and workload management. However, it is difficult to extend for specific workflows of other communities. The DIRAC architecture and development framework is conceived to have excellent potential for extension of its functionality. Therefore, completeness of its base functions together with modular extendable architecture makes DIRAC a unique all-in-one solution suitable for many scientific applications.

In this paper, we review the DIRAC Project giving details about its general architectures as well as about workload and data management capabilities in Section 2. Examples of the system usage are described in Section 3 followed by Conclusions.

2 DIRAC Overview

DIRAC Project provides all the necessary components to create and maintain distributed computing systems. It forms a layer on top of third party computing infrastructures, which isolates users from the direct access to the computing resources and provides them with an abstract interface hiding the complexity of dealing with multiple heterogeneous services. This pattern is applied to both computing and storage resources. In both cases, abstract interfaces are defined and implementations for all the common computing service and storage technologies are provided. Therefore, users see only logical computing and storage elements, which simplifies dramatically their usage. In this section, we will describe in more details the DIRAC systems for workload and data management.

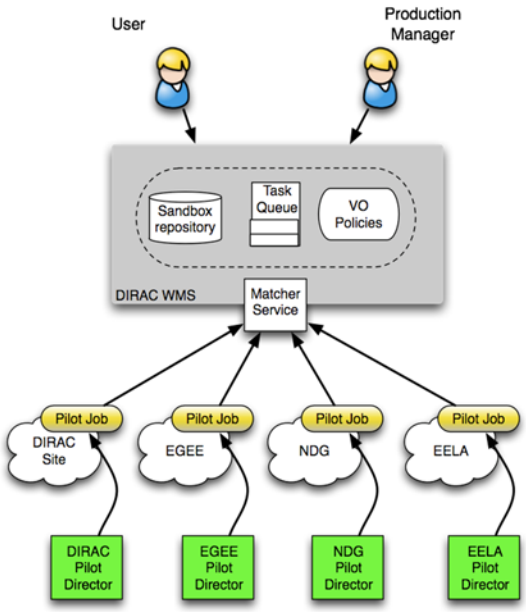


Figure 1 WMS with pilot jobs

2.1 Workload Management

The DIRAC Workload Management System is based on the concept of pilot jobs [2]. In this scheduling architecture (Figure 1), the user tasks are submitted to the central Task Queue service. At the same time, the so-called pilot jobs are submitted to the computing resources by specialized components called Directors. Directors use the job scheduling mechanism suitable for their respective computing infrastructure: grid resource brokers or computing elements, batch system schedulers, cloud managers, etc. The pilot jobs start execution on the worker nodes, check the execution environment, collect the worker node characteristics and present them to the Matcher service. The Matcher service chooses the most appropriate user job waiting in the Task Queue and hands it over to the pilot for execution. Once the user task is executed and its outputs are delivered to the DIRAC central services, the pilot job can take another user task if the remaining time of the worker node reservation is sufficient.

There are many advantages of the pilot job concept. The pilots are not only increasing the visible efficiency of the user jobs but also help managing heterogeneous computing resources presenting them to the central services in a uniform coherent way. Large user communities can benefit also from the ability of applying the community policies that are not easy, if at all

possible, with the standard grid middleware. Furthermore executing several user tasks in the same pilot largely reduces the stress on the batch systems no matter if they are accessed directly or via grid mechanisms, especially if users subdivide their payload in many short tasks trying to reduce the response time.

The pilot job based scheduling system allows easy aggregation of computing resources of different technologies. Currently the following resources are available for DIRAC users:

- Computing grid infrastructures based on the gLite/EMI grid middleware. The submission is possible both through the gLite Workload Management System and directly to the computing element services exposing the CREAM interface. WLCG and EGI grids are examples of such grid infrastructures.
- Open Science Grid (OSG) infrastructure based on the VDT (Virtual Data Toolkit) suite of middleware [3].
- Grids based on the ARC middleware, which was developed in the framework of the Nordugrid project [4].
- Standalone computing clusters with common batch system schedulers, for example, PBS/Torque, Grid Engine, Condor, SLURM, OAR, and others. Those clusters can be accessed by configuring an SSH tunnel to be used by DIRAC directors to submit pilot jobs to the local batch systems. No specific services are needed on such sites to include them into a distributed computing infrastructure.
- Sites providing resources via most widely used cloud managers, for example OpenStack, OpenNebula, Amazon and others. Both commercial and public clouds can be accessed through DIRAC.
- Volunteer resources provided with the help of BOINC software. There are several realizations of access to this kind of resources all based on the same pilot job framework.

As it was explained above, a new kind of computing resources can be integrated into the DIRAC Workload Management System by providing a corresponding Director using an appropriate job submission protocol. This is the plugin mechanism that enables easily new computing facilities as needed by the DIRAC users.

2.2 Data Management

The DIRAC Data Management System (DMS) is based on similar design principles as the WMS [5]. An abstract interface is defined to describe access to a storage system with multiple implementations for various storage access protocols. Similarly, there is a concept of a FileCatalog service, which provides information about the physical locations of file copies. As for storage services there are several implementations for different catalog service technologies all following the same abstract interface.

A particular storage system can be accessible via different interfaces with different access protocols. But for the users it stays logically a single service providing access to the same physical storage space. To simplify access to this kind of services, DIRAC aggregates plugins for different access protocols according to the storage service configuration description. When accessing the service, the most appropriate plugin is chosen automatically according to the client environment, security requirements, closeness to the service, etc. As a result, users are only seeing logical entities without the need to know the exact type and technology of the external services.

DIRAC provides plug-ins for a number of storage access protocols most commonly used in the distributed storage services:

- SRM, XRootd, RFIO, etc;
- gfal2 library based access protocols (DCAP, HTTP-based protocols, S3, WebDAV, etc) [6].

New plug-ins can be easily added to accommodate new storage technologies as needed by user communities.

In addition DIRAC provides its own implementation of a Storage Element service and the corresponding plug-in using the custom DIPS protocol. This is the protocol used to exchange data between the DIRAC components. The DIRAC StorageElement service allows exposing data stored on file servers with POSIX compliant file systems. This service helps to quickly incorporate data accumulated by scientific communities in any *ad hoc* way into any distributed system under the DIRAC interware control.

Similarly to Storage Elements, DIRAC provides access to file catalogs via client plug-ins. The only plug-in to an external catalog service is for the LCG File Catalog (LFC), which used to be a *de facto* standard catalog in the WLCG and other grid infrastructures. Other available catalog plug-ins are used to access the DIRAC File Catalog (DFC) service and other services that are written within the DIRAC framework and implement the same abstract File Catalog interface [7]. These plug-ins can be aggregated together so that all the connected catalogs are receiving the same messages on new data registration, file status changes, etc. The usefulness of aggregating several catalogs can be illustrated by an example of a user community that wants to migrate the contents of their LFC catalog to the DFC catalog. In this case, the catalog data can be present in both catalogs for the time of migration or for redundancy purpose.

The DIRAC File Catalog has the following main features:

- Standard file catalog functionality for storing file metadata, ACL information, checksums, etc.
- Complete Replica Catalog functionality to keep track of physical copies of the files.
- Additional file metadata to define ancestor-descendent relations between files often needed for applications with complex workflows.

- Efficient storage usage reports to allow implementation of community policies, user quotas, etc.
- Metadata Catalog functionality to define arbitrary user metadata on directory and file levels with efficient search engine.
- Support for dataset definition and operations.

The DFC implementation is optimized for efficient bulk queries where the information for large numbers of files is requested in case of massive data management operations. Altogether, the DFC provides logical name space for the data and, together with storage access plug-ins, makes data access as simple as in a distributed file system.

Storage Element and File Catalog services are used to perform all the basic operations with data. However, bulk data operations need special support so that they can be performed asynchronously without a need for a user to wait for the operation completion at the interactive prompt. DIRAC Request Management System (RMS) provides support for such asynchronous operations. Many data management tasks in large scientific communities are often repeated for different data sets. DIRAC provides support for automation of recurrent massive data operations driven by the data registration or file status change events. Other data related services include:

In addition to the main DMS software stack, DIRAC provides several more services helping to perform particular data management tasks:

- Staging service to manage bringing data on-line into a disk cache in the SEs with tertiary storage architecture;
- Data Logging service to log all the operations on a predefined subset of data mostly for debugging purposes;
- Data Integrity service to record failures of the data management operations in order to spot malfunctioning components and resolve issues;
- The general DIRAC Accounting service is used to store the historical data of all the data transfers, success rates of the transfer operations.

2.3 DIRAC development framework

All the DIRAC components are written in a well-defined software framework with a clear architecture and development conventions. Since large part of the functionality is implemented as plug-ins implementing predefined abstract interfaces, extending DIRAC software to cover new cases is simplified by the design of the system. There are several core services to orchestrate the work of the whole DIRAC distributed system, the most important ones are the following:

- Configuration service used for discovery of the DIRAC components and providing a single source of configuration information;
- Monitoring service to follow the system load and activities;
- Accounting service to keep track of the resources consumption by different communities, groups and individual users;
- System Logging service to accumulate error reports in one place to allow quick reaction to occuring problem.

Modular architecture and the use of core services allow developers to easily write new extensions concentrating on their specific functionality and avoiding recurrent tasks.

All the communications between distributed DIRAC components are secure following the standards introduced by computational grids, which is extremely important in the distributed computing environment. A number of interfaces are provided to users to interact with the system. This includes a rich set of command-line tools for Unix environment, Python language API to write one's own scripts and applications, RESTful interface to help integration with third party applications. DIRAC functionality is available also through a flexible and secure Web Portal which follows the user interface paradigm of a desktop computer.

3 DIRAC Users

DIRAC Project was initiated by the LHCb experiment at CERN. LHCb stays the most active user of the DIRAC software. The experiment data production system ensures a constant flow of jobs of different kinds: reconstruction of events of proton-proton collisions in the LHC collider, modelling of the LHCb detector response to different kinds of events, final user analysis of the data [8]. Figure 2. illustrates the scale of computing resources usage by the LHCb experiment. As it can be seen, there are on average about 50 thousands jobs running simultaneously on more than 120 sites, with peak values going to up to 100 thousands jobs. This is equivalent to operating a virtual computing centre of about 100 thousands of processor cores. At the same time the total data volume of LHCb exceeds 10 PB distributed over more than twenty millions of files, many of those having 2 and more physical copies in about 20 distributed storage systems. Information about all these data is stored in the DIRAC File Catalog. LHCb has created a large number of extensions to the core DIRAC functionality in order to support its specific workflows. All these extensions are implemented in the DIRAC development framework and can be released, deployed and maintained using standard DIRAC tools.

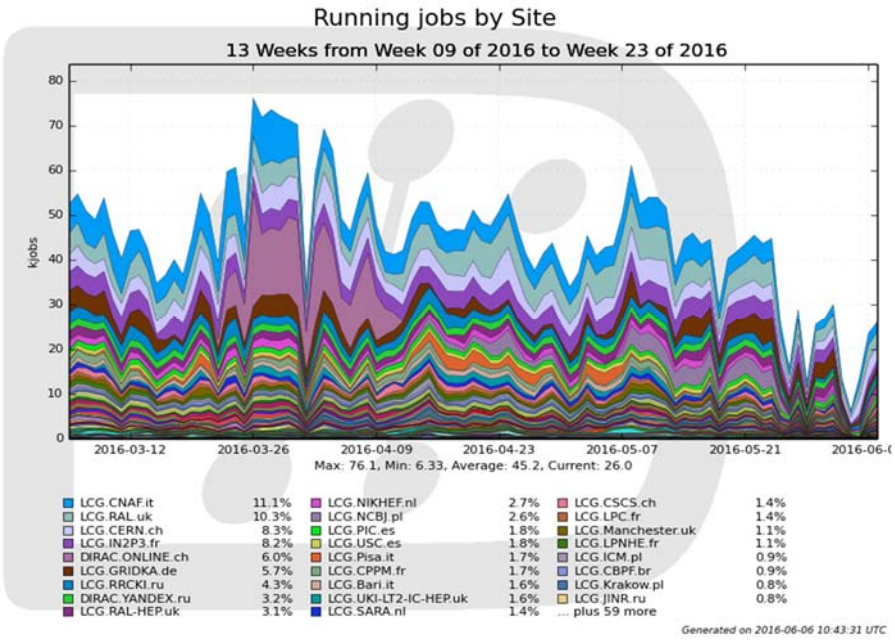


Figure 2 Running jobs of the LHCb experiment

After the DIRAC

system was successfully used within LHCb, several other experiments in High Energy Physics and other domains expressed interest in using this software for their data production systems, for example: BES III experiment at the BEPC collider in Beijing, China [9]; Belle II experiment at the KEK centre, Tsukuba, Japan [10]; the CTA astrophysics experiment being constructed now in Chile [11], and others. Open architecture of the DIRAC project was easy to adapt for the workflows of particular experiments. All of them developed several extensions to accommodate their specific requirements all relying on the use of the common core DIRAC services.

3.1 DIRAC as a service

Experience accumulated by running data intensive applications of the High Energy Physics experiments can be very valuable for researchers in other scientific domains, which have high computing requirements. However, if even the DIRAC client software is easy to install and use, running dedicated DIRAC services requires a high expertise level and is not easy especially for the research communities without long-term experience in large-scale computations. Therefore, several national computing infrastructure projects are offering now DIRAC services for their users. The first example of such service was created by the France-Grilles National Grid Initiative (NGI) project in France [12].

By 2011 in France, there were several DIRAC service installations used by different scientific or regional communities. There was also a DIRAC service maintained by the France-Grilles NGI as part of its training and dissemination program. This allowed several teams of experts in different universities to gain experience with installation and operation of DIRAC services. However, the combined maintenance effort for multiple DIRAC service instances was quite high. Therefore, it was proposed to integrate independent DIRAC installations into a single national service to optimize operational costs. The responsibilities of different partners of the project were distributed as follows. The France-Grilles NGI (FG) ensures the overall coordination of the project. The IN2P3 Computing Centre (CC/IN2P3) hosts the service providing the necessary hardware and manpower. The service is operated by a distributed team of experts from several laboratories and universities participating to the project.

From the start, the FG-DIRAC service was conceived for usage by multiple user communities. Now it is intensively used by researchers in the domains of life sciences, biomedicine, complex system analysis, and many others. It is very important that user support and assistance in porting applications to the distributed computing environments is the integral part of the service. This is especially needed for research domains where the computing expertise is historically low. Therefore, the France-Grilles NGI organizes multiple tutorials for interested users based on the FG-DIRAC platform. The tutorials not only demonstrate basic services capabilities but are also used to examine cases of particular applications and the necessary steps to start running them on distributed computing resources. The service has an active user community built around it and provides a forum where researchers are sharing their experience and helping the newcomers.

After the successful demonstration of DIRAC services provided by the French national computing research infrastructure, similar services were deployed in

some other countries: Spain, UK, China, and some others. There are several ongoing evaluation projects testing the DIRAC functionality and usability for similar purposes. Since 2014, DIRAC services are provided by the European Grid Initiative (EGI) for the research communities in Europe and beyond [13].

A general-purpose DIRAC service was deployed in the Laboratory of Information Technologies in JINR, Dubna. This service is provided to users participating in international collaborations that already use DIRAC tools. It is also used for evaluation of DIRAC as a distributed computing platform for experiments that are now under preparation in JINR. The service is providing access to computing resources of the WLCG and EGI grid infrastructures. It has also several High Performance Computing (HPC) centers connected and offers a possibility to create complex workflows including massively parallel applications. The service is planned to become a central point for a federation of HPC centers in Russia and other countries. It will provide a framework for unified access to the HPC centers similar to existing grid infrastructures.

4 Conclusions

DIRAC interware is a versatile software suite for building distributed computing systems. It has gone a long way of development starting from a specific tool for a large-scale High Energy Physics experiment and is now available as a general-purpose product. Various computing and storage resources based on different technologies can be incorporated under the overall control by the DIRAC Workload and Data Management Systems. The open architecture of the DIRAC software allows easy connection of the new emerging types of resources as needed by the user communities. The system is designed for extensibility to support specific workflows and data requirements of particular applications. Completeness of its functionality as well as its modular design can ensure solution for a variety of distributed computing tasks and for a wide range of scientific communities in a single framework.

The number of DIRAC users is growing with the applications coming from various scientific domains. A number of multi-community DIRAC services are provided now by several national computing infrastructure projects are available to support small research communities not having dedicated systems for managing distributed computing resources. This helps many researchers without a deep special computing expertise level to get familiar with using distributed computing systems by following specialized tutorials and benefitting from assistance in porting their applications to this environment. Altogether, this makes large-scale data intensive computations more accessible improving the overall quality of their scientific results.

References

[1] A. Tsaregorodtsev et al, DIRAC3 : The New Generation of the LHCb Grid Software, 2010 J. Phys.: Conf. Ser., 219 062029; DIRAC Project - http://diracgrid.org

[2] A.Casajus, R.Graciani, A.Tsaregorodtsev, DIRAC pilot framework and the DIRAC Workload Management System, 2010 J. Phys.: Conf. Ser. 219 062049

[3] OpenScience Grid - https://www.opensciencegrid.org/

[4] ARC project - http://www.nordugrid.org/arc/

[5] A.Smith, A.Tsaregorodtsev, DIRAC: data production management, 2008 J. Phys.: Conf.Ser. 119 062046

[6] Gfal2 Project - https://dmc.web.cern.ch/projects-tags/gfal-2

[7] S. Poss and A. Tsaregorodtsev, DIRAC File Replica and Metadata Catalog, 2012 J. Phys.: Conf.Ser. 396 032108

[8] F. Stagni F and Ph. Charpentier, The LHCb DIRAC-based production and data management operations systems, 2012 J. Phys.: Conf. Ser. 368 012010

[9] X.M. Zhang, I. Pelevanyuk, V. Korenkov et al, Design and Operation of the BES-III Distributed Computing System, 2015 Procedia Computer Science 66

[10] T.Kuhr, T.Hara, Computing at Belle II, 2015 J. Phys.: Conf. Ser. 396 032063

[11] L.Arrabito et al, Application of the DIRAC framework in CTA: first evaluation, 2015 J. Phys.: Conf. Ser. 396 032007

[12] France-Grilles DIRAC portal – http://dirac.france-grilles.in2p3.fr

[13] DIRAC4EGI service portal – http://dirac.egi.eu

[14] A.Klimentov et al, Next Generation Workload Management System For Big Data on Heterogeneous Distributed Computing, 2015 J. Phys.: Conf. Ser. 608 012040

[15] S. Bagnasco, L. Betev, P. Buncic et al, AliEn: ALICE environment on the GRID, 2008 J. Phys.: Conf. Ser. 119 062012

Поиск компонентов источников гравитационных волн в электромагнитном диапазоне и с помощью методов астрономии космических лучей

© А. С. Позаненко © А. А. Вольнова © П. Ю. Минаев

Институт космических исследований Российской академии наук, Москва
apozanen@iki.rssi.ru alinusss@gmail.com minaevp@mail.ru
© В. А. Самодуров

Пущинская радиоастрономическая обсерватория, Пущино
Национальный исследовательский университет «Высшая школа экономики», Москва
sam@prao.ru

Аннотация

Рассмотрены примеры возможных сигналов, сопровождающих источники гравитационных волн, обсуждаются проблемы поиска этих сигналов в различных диапазонах электромагнитного излучения, наборы данных для такого поиска, типы данных, на которых производится поиск, и сформулированы требования к поиску. В качестве примера рассматриваются впервые зарегистрированные в эксперименте LIGO гравитационно-волновые события GW150914, GW151226 и LVT151012.

1 Введение

Одной из актуальных задач современной астрофизики является поиск компонентов всплесков гравитационного излучения. В 2015 г. исполнилось 100 лет с тех пор, как Альберт Эйнштейн сформулировал основные принципы Общей Теории Относительности (ОТО) [1]. Большинство наблюдаемых следствий ОТО нашло подтверждение прямыми и многократными наблюдениями. И вот пришел черед подтверждению существования гравитационных волн. 14 сентября 2015 года гравитационные волны (ГВ) были обнаружены двумя пространственно разнесенными детекторами эксперимента LIGO [2]. Насколько надежно открытие? Это главный вопрос экспериментальной физики. История науки знает массу ошибок, сделанных из-за неправильной оценки достоверности. В данном случае совместная значимость того, что оба детектора зарегистрировали не случайное событие, а реальный сигнал, составляет более 5 стандартных отклонений. Несмотря на то, что значимость регистрации сигнала достаточно велика, необходимы дальнейшие подтверждения. И они были получены последующим детектированием еще двух событий (LVT151012 и GW151226) в октябре и декабре 2015 г. [19]. Большая часть научного сообщества считает, что трех регистраций достаточно для подтверждения обнаружения гравитационных волн. Однако есть и немалая группа ученых, сомневающихся в надежности открытия. Совместная регистрации гравитационного события с событием в электромагнитном диапазоне (гамма-излучение, оптика, радио) и/или путем регистрации нейтрино от источника гравитационных волн могла бы положить конец этой дискуссии. Такие наблюдения позволят однозначно связать источник ГВ с уже известными источниками в других диапазонах. Кроме отождествления источника ГВ это открывает новую страницу гравитационной астрономии, когда совместное изучение наблюдательных данных разного типа взаимодействий даст возможность заглянуть туда, куда нельзя заглянуть ни с помощью наблюдений в электромагнитном диапазоне, ни даже с помощью нейтринной астрономии, т.е. заглянуть в такие состояния вещества, которые могут существовать только лишь при коллапсе массивных звезд или же при слиянии тесных релятивистских двойных систем, когда ни электромагнитное излучение, ни даже нейтрино не могут «выбраться» из сверхплотных состояний вещества.

Для дальнейшего исследования ГВ требуется увеличение количества их регистраций и, конечно, необходимо искать источники, сопровождающие излучение ГВ в оптическом или ином диапазоне электромагнитного излучения. Мы рассмотрим примеры возможных наблюдаемых источников, сопровождающих ГВ, проблемы поиска в различных спектрах электромагнитного излучения, наборы

Труды XVIII Международной конференции DAMDID/RCDL'2016 «Аналитика и управление данными в областях с интенсивным использованием данных», Ершово, 11-14 октября 2016

данных для такого поиска, типы данных, на которых производится поиск, и сформулируем требования к поиску. Предложение конкретных решений находится вне рамок нашей работы.

2 Источники и сигналы от них

Источниками ГВ, доступными для детектирования в настоящее время, являются взрывы Сверхновых с коллапсирующим ядром в ближайших галактиках и слияния тесных двойных или же кратных систем, в которых есть пара в любом сочетании: нейтронная звезда (НЗ)/черная дыра (ЧД). Первое зарегистрированное гравитационно-волновое событие GW150914 было слиянием системы двух ЧД на расстоянии от Земли, эквивалентном космологическому красному смещению z=0.09 [2]. Аналогично, два последующих события, GW151226 (значимость 5.3 стандартных отклонения), и LVT151012 (1.7 стандартных отклонения) также представляют собой слияние двойных систем, состоящих из черных дыр и расположенных на космологических расстояниях z=0.09 и z=0.20 [19]. Так как значимость LVT151012 невелика, то это событие считается кандидатом, что, впрочем, не мешает включать его в статистический анализ распределения черных дыр по массам, распределения ЧД и частоты слияния ЧД в локальной Вселенной и т.п. [19]. За время цикла наблюдений с 12 сентября 2015 по 19 января 2016 не было зарегистрировано гравитационно-волновых событий от слияния систем типа НЗ/НЗ или НЗ/ЧД, это позволило получить верхний предел на частоту слияния таких систем в локальном объеме Вселенной, и предел оказался существенно меньше, чем частота слияний ЧД/ЧД [20]. Это не является оптимистичным прогнозом для поиска компонентов гравитационно-волновых событий в спектре электромагнитного излучения.

Сигнал от слияния двойной системы смоделирован и представляет собой пакет из увеличивающихся по амплитуде колебаний с уменьшающимся периодом. Такой «рисунок» (шаблон) и ищется вейвлетным анализом во временных рядах величин смещений пробных тел в детекторах LIGO. Локализация же проводится методом триангуляции, где задержка прихода сигнала на один детектор относительно других детекторов (в нестоящее время работают два детектора) определят область допустимых значений положения источника на небесной сфере. Если время регистрации известно достаточно точно и определяется, в основном, лишь интенсивностью зарегистрированного сигнала, то область локализации определяется существенно менее точно и может составлять сотни квадратных градусов.

Сверхновые с коллапсирующим ядром (SN II, SN Ib/c) проявляют себя в оптическом диапазоне давно и успешно наблюдаются как быстро нарастающий по потоку точечный объект, расположенный в родительской галактике. Максимума кривая блеска достигает через несколько дней после начала роста и затем происходит более медленный спад. В некоторых редких случаях Сверхновая сопровождается классическим (длинным) гамма-всплеском. Впрочем, обычно говорят, что гамма-всплеск сопровождается появлением сверхновой. Излучение гамма-всплеска регистрируется как в гамма-диапазоне, так и в фазе послесвечения, в оптическом, рентгеновском и радио- диапазонах [3].

Электромагнитное излучение может возникнуть и при слиянии систем НЗ/НЗ, НЗ/ЧД в виде короткого гамма-всплеска длительностью существенно меньше, чем в предыдущем случае, а также менее интенсивным (по сравнению с длинными всплесками) послесвечением также регистрируемом в оптическом, рентгеновском и радио- диапазонах. В течение нескольких суток после слияния может наблюдаться «килоновая». Во многом килоновая аналогична понятию сверхновой и связана с распадом радиоактивных элементов, синтезированных при слиянии двойной системы [21]. Килоновая уже наблюдалась, по крайней мере, в одном случае - в кривой блеска послесвечения короткого гамма-всплеска GRB 130603B [22].

А вот слияние системы ЧД/ЧД не должно сопровождаться сколько-нибудь значимым электромагнитным излучением. Впрочем, совсем близко, через 0.4 сек после регистрации гравитационно-волнового события GW150914 гамма-детекторы космического эксперимента GBM на обсерватории Ферми зарегистрировали сигнал, морфологически похожий именно на короткий гамма-всплеск [4]. Хотя это может быть и простым совпадением во времени. Собственно, точность локализации на небесной сфере, а также точность совпадения во времени и являются главными проблемами при архивном поиске, а скорость реакции на ГВ, полнота покрытия и глубина обзора – при поиске в режиме реального времени.

3 Проблемы поиска

Точность локализации источника гравитационных волн на небесной сфере составляет сотни квадратных градусов. Апертурные телескопы в радио-, оптическом, рентгеновском и гамма-диапазонах не способны охватить такое большое поле зрения за одно наведение. Необходимы скоординированные наблюдения многих обсерваторий на Земле для сканирования всей области локализации. Если сигнал в оптическом или радио- диапазоне, сопровождающий гравитационно-волновое событие, достаточно продолжительный, то при сканировании действительно можно обнаружить транзиентный объект. Но таких транзиентных объектов различной природы в достаточно большой области локализации может быть несколько, и, следовательно, предстоит выбирать, какой именно из транзиентных объектов принадлежит гравитационно-волновому событию.

Всенаправленные детекторы гамма-диапазона, расположенные на орбите Земли на космических обсерваториях (например, GBM/Fermi, Konus-Wind, SPI-ACS/INTEGRAL, БДРГ/Ломоносов) способны регистрировать сигналы одновременно со всего неба, но не способны производить точную локализацию. В этом случае искать нужный сигнал можно по совпадению во времени. Однако и тут остается фактор возможного случайного совпадения. Вопрос детектирования компонента гравитационно-волнового события в электромагнитном диапазоне может быть решен, например, или путем синхронной регистрации в гамма- и оптическом или радио-диапазонах, или регистрации в гамма-диапазоне с последующим отождествлением источника в оптике и радио, аналогично тому, как сейчас происходит отождествление компонентов космических гамма-всплесков. Во многом, задача идентификации гравитационно-волнового события подобна тому, что происходило с поиском компонентов гамма-всплесков с момента их открытия (1967) [5] до первого обнаружения в рентгеновском и оптическом диапазоне (1998) [6]. Первая широкомасштабная кампания поиска была проведена, начиная, примерно, с суток после регистрации GW 150914 во всех диапазонах и длилась в течение нескольких месяцев [7-10]. Результатов, кроме уже упоминавшегося короткого гамма-всплеска [4], кампания не принесла.

Разнородность данных препятствует строго одинаковому поиску даже в данных одного типа. Например, разные энергетические диапазоны гамма-детекторов, наблюдения в различных широкополосных фильтрах разных областей возможной локализации и т.п.

Еще одной неопределенностью можно считать незнание всех возможных форм сигнала в электромагнитном диапазоне, сопровождающих, и возможно предшествующих, гравитационно-волновому событию. Кроме уже рассмотренных выше, необходимо искать пакеты периодических или квазипериодических осцилляций в гамма-диапазоне, связанных с колебаниями аккреционного диска, возникающего после слияния компактных систем. Впрочем, так как мы не знаем деталей слияния, то возможна регистрация и нового типа сигналов, не подпадающих ни под один рассмотренный шаблон.

4 Наборы и типы данных, инфраструкутра

Достаточно разнородные данные соответствуют различным диапазонам, но везде - это в том или ином виде временные ряды. В гамма-диапазоне - это временные ряды или равноотстоящих по времени бинов или данные записи каждого фотона, зарегистрированного всенаправленным детектором (время регистрации фотона и его энергия). Наиболее подходящие для такого поиска работающие в настоящее время эксперименты GBM/Fermi, Konus-Wind, SPI-ACS/INTEGRAL. Данные экспериментов GBM/Fermi, SPI-ACS/INTEGRAL являются публично доступными. Во всех трех экспериментах, а также множестве других, предназначенных для изучения космических гамма-всплесков, существует система автоматического выделения всплеска, основанная на поиске кратковременного превышении сигнала над медленно меняющимся фоном. Но не всегда значимость того или иного транзиентного события достаточна для автоматического отбора, и необходимо использовать архивы данных для повторного поиска транзиентов, соответствующих гравитационно-волновому событию. Именно так и был найден возможный компонент GW 150914 в эксперименте GBM/Fermi [4]. В данных других аналогичных экспериментов (Konus-Wind, SPI-ACS) этот сигнал найден не был, наиболее вероятно, из-за более низкой чувствительности.

Данные нейтринных обсерваторий и данные обсерваторий космических лучей (высокоэнергичных фотонов и заряженных частиц с энергией ~1 ГэВ и более) также надо рассматривать при поиске компонента гравитационно-волнового события. Эти данные аналогичны пофотонной записи, рассмотренной выше, но для космических лучей и для высокоэнергичного мюонного нейтрино можно определить еще и положение возможного источника на небесной сфере, впрочем, это положение также весьма приблизительное. Нейтринный телескоп IceCube [11], отечественные нейтринные телескопы Байкальской нейтринной и Баксанской нейтринной обсерватории [12] необходимо использовать при поиске компонентов. Эксперимент IceCube уже работает как по программе поиска совпадений, так и сам является источником для организаций кампаний, аналогичных гравитационно-волновым, а именно по поиску транзиентов, сопровождающих штучно зарегистрированные нейтрино [13].

Данные апертурных телескопов представляют собой те же самые временные ряды, но только записанные с различных участков небесной сферы (оптические телескопы) и иногда в нескольких различных диапазонах длин волн (радиотелескопы). Среди радиотелескопов существуют телескопы с фазированным сбором сигнала с множества антенн, которые позволяют осматривать одновременно большие участки неба (БСА, LOFAR). Данные с таких телескопов наиболее пригодны к поиску плохо локализованных источников. Например, данные БСА уже позволили открыть несколько новых пульсаров [14,15], а данные LOFAR были использованы для обзора части неба, совпадающей с северной частью локализации GW 150914.

Оптические данные, возможно, самые информативные, как с точки зрения полезности информации, так и с точки зрения информационного объема. Дело в том, что в оптическом диапазоне точность локализации максимальна, и это позволяет идентифицировать найденный объект с уже имеющимся в каталогах (например, с галактикой)

и/или целенаправленно исследовать свойства объекта с помощью дополнительных, более чувствительных наблюдений. Наиболее полезны обзоры, позволяющие отсканировать всю область локализации одним и тем же инструментом, но зачастую это принципиально невозможно, т.к. исходная область локализации может находиться и в северной, и в южной части неба, как это и было с источником GW 150914. Системы существующих телескопов iPTF, Pan-STARRS, MASTER и планирующихся к введению в строй, AZT-33BM, LSST наиболее пригодны к такому типу обзоров.

Оптические данные представляют собой данные с ПЗС-матрицы, т.е. двумерную матрицу значений сигнала, записанную с NxM пикселей ПЗС-матрицы. Как правило, исследования проводят не с исходными данными (как в гамма-диапазоне и иногда в радиодиапазоне), а с уже редуцированными данными и с построенными на их основе каталогами, т.е. с обнаруженными на снимках объектами, выделенными специальной процедурой и имеющие, по крайней мере, координаты и оценку потока [16]. Исследования заключаются в сравнении потока от известных ранее объектов (например, из других каталогов, или же из каталогов своих же, более ранних наблюдений) и/или поиске новых объектов, имеющих признак предполагаемого явления (кратковременная вспышка, сверхновая, килоновая, послесвечение гамма-всплеска, или же так называемый гамма-всплеск-сирота, т.е. гамма-всплеск, не сопровождающийся вспышкой именно в гамма-диапазоне). Именно редукция и составление каталогов являются наиболее времязатратной процедурой [17].

5 Требования к поиску

Коротко рассмотрим требования к наблюдениям и поиску сигналов, сопровождающих гравитационно-волновые событие. Можно коротко сформулировать их так: быстро, глубоко, во всех диапазонах, с полным охватом возможной области локализации.

При выборе глубины (чувствительности) обзора необходимо максимизировать глубину охвата, т.е. объем пространства, просматриваемый при обзоре. Иногда этот показатель называют термином «grasp».

Очевидно, что время начала обзора должно быть минимизировано, также как и время оповещения о необходимости таких наблюдений. Широта охвата возможных наблюдателей должна быть максимальной. Последнее, кстати, не выполняется в консорциуме LIGO/Virgo, где подписка на оповещения возможна лишь после подписания соответствующего соглашения с консорциумом.

Чрезвычайно важной является скорость обработки полученных данных, т.е. выделения новых источников и сравнения с существующими каталогами. Идеальной скоростью была бы обработка в режиме реального времени.

Важным является проведение наблюдений во всех диапазонах. В экспериментах с использованием космических гамма-детекторов это происходит автоматически (если они находятся во включенном состоянии). В обзорных радиотелескопах (LOFAR, БСА), наблюдающих одновременно большую часть небесной сферы, наблюдения возможной области локализации также происходят автоматически при совпадении части области локализации с полем зрения телескопа.

При поиске важно использовать все различные модели предполагаемого сигнала как для оптимизации поиска, так и для увеличения чувствительности поиска на представленных данных. Например, если происходит поиск периодического, но ограниченного во времени сигнала, то целесообразно использовать вейвлетный анализ с базовыми квазипериодическими функциями.

Необходимо проводить поиск в архивных данных, наблюдения которых могут соответствовать времени появления гравитационно-волнового события или же покрывать область его локализации.

Поскольку данные неоднородны, так как генерируются различными инструментами, то необходимо проведение взаимных калибровок для совместной оценки значимости найденных сигналов, а в случае их отсутствия, совместной оценки пределов на потоки излучения от предполагаемых источников. И, конечно же, необходима инфраструктура для быстрого и наиболее широкого распространения полученных результатов об обнаружении или же не обнаружении источников в области локализации гравитационно-волнового сигнала. В нестоящее время эту роль успешно выполняет сеть GCN [18].

6 Вместо заключения

Поиск компонентов гравитационно-волновых событий уже начался, и в немалой степени успешность поиска будет зависеть от координированного информационного обеспечения поисковых наблюдений, эффективных алгоритмов обработки и их программной реализации.

Благодарности

Работа частично поддержана грантами РФФИ 16-07-01028 и 16-32-00489 мол_а.

Литература

[1] Einstein, 1915, «Die Feldgleichungen der Gravitation». Sitzungsberichte der Preussischen Akademie der Wissenschaften zu Berlin: 844—847.

[2] B.P. Abbot *et al*., 2016, Phys.Rev.Lett., 116, 24, id.241103

[3] N. Gehrels, E. Ramirez-Ruiz, and D.B. Fox, 2009, Annual Review of Astronomy and Astrophysics, Vol. 47: 567-617

[4] V. Connaughton, E. Burns, A. Goldstein, *et al.*, 2016, arXiv:1602.03920.

[5] R.W. Klebesadel, I.B. Strong, and R.A. Olson 1973, Ap.J. (Letters) 182, L85.

[6] J. Castro-Tirado, J. Gorosabel, N. Benitez, *et al.*, 1998, Science, Vol. 279, Iss. 5353, p. 1011.

[7] T. Morokuma, M. Tanaka, Y. Asakura, *et al.*, 2016, eprint arXiv:1605.03216.

[8] Z. Bagoly, D. Szécsi, L. G. Balázs, *et al.*, 2016, eprint arXiv:1603.06611.

[9] M. M. Kasliwal, S. B. Cenko, L. P. Singer, *et al.*, 2016, eprint arXiv:1602.08764.

[10] S. J. Smartt, K. C. Chambers, K. W. Smith, *et al.*, 2016, eprint arXiv:1602.04156.

[11] IceCube Neutrino Observatory https://icecube.wisc.edu/

[12] Баксанская нейтринная обсерватория (БНО) ИЯИ РАН http://www.inr.ru/bno.html

[13] S. Adrián-Martínez, A. Albert, M. André, *et al.*, 2016, The Astrophysical Journal, v. 823, article id. 65.

[14] S. A. Tyul'bashev, V. S. Tyul'bashev, V. V. Oreshko, S. V. Logvinenko, 2016, Astronomy Reports, Volume 60, Issue 2, pp.220-232.

[15] V. A. Samodurov, A. E. Rodin, M. A. Kitaeva, E. Isaev, D. V. Dumskij, D. Churakov, M. Manzyuk The daily 110MHz sky survey (BSA FIAN): online database, science goals data processing by distributed computing. Труды XVII международной конференции «Аналитика и управление данными в областях с интенсивным использованием данных» (“Data Analytics and Management in Data Intensive Domains” (DAMDID)). Обнинск: НИЯУ МИФИ, 2015. P. 127-128.

[16] А.С. Позаненко, А.А. Вольнова, Использование астрономических каталогов для поиска аномальных объектов. Аналитика и управление данными в областях с интенсивным использованием данных: XVII Международная конференция DAMDID / RCDL’2015 (Обнинск, 13-16 октября 2015 года, Россия): Труды конференции / под ред. Л.А. Калиниченко, С.О. Старкова – Обнинск, ИАТЭ НИЯУ МИФИ, С.315, 2015.

[17] Л. А. Калиниченко, Е. П. Гордов, А. А. Вольнова, Н. Н. Киселева, Д. А. Ковалева, О. Ю. Малков, И. Г. Окладников, Н. Л. Подколодный, А. С. Позаненко, Н. В. Пономарева, С. А. Ступников, А. З. Фазлиев Проблемы доступа к данным в исследованиях с интенсивным использованием данных в России. Аналитика и управление данными в областях с интенсивным использованием данных: XVII Международная конференция DAMDID / RCDL’2015 (Обнинск, 13-16 октября 2015 года, Россия): Труды конференции / под ред. Л.А. Калиниченко, С.О. Старкова – Обнинск, ИАТЭ НИЯУ МИФИ, С.387, 2015.

[18] http://gcn.gsfc.nasa.gov/gcn3_archive.html

[19] P. Abbott, *et al*., 2016, eprint arXiv:1606.04856

[20] P. Abbott, *et al*., 2016, eprint arXiv:1607.07456

[21] L. Li, B. Paczynski, 1998, ApJ, 507, L59

[22] N. R. Tanvir, A. J. Levan, A. S. Fruchter, *et al*., 2013, Nature, 500, 547

A search of counterparts of the sources of gravitational waves in different wavelength of electromagnetic radiation and with methods of cosmic rays astronomy

Alexei S. Pozanenko, Alina A. Volnova, Pavel Yu. Minaev, Vladimir A. Samodurov

We consider various signals, which may accompany transient sources of gravitational waves. Also we discuss problems of a counterpart search of gravitational transients. We briefly describe data sets and data types generating in search campaigns. We also formulate requirements for the search. As an example we consider the first gravitational wave events GW150914, GW151226 and candidate LVT151012 detected by the LIGO experiment.

Исследовательские инфраструктуры в материаловедении

Research Data Infrastructures in Material Sciences

Инфраструктура обеспечения данными специалистов в неорганической химии и материаловедении

© Н. Н. Киселева[1] © В. А. Дударев[1,2]

[1]Федеральное государственное бюджетное учреждение науки Институт металлургии и материаловедения им. А.А. Байкова Российской академии наук (ИМЕТ РАН),
[2]Национальный исследовательский университет «Высшая школа экономики» (НИУ ВШЭ), Москва

kis@imet.ac.ru vic@imet.ac.ru

Аннотация

Проведен анализ реализуемых в мире крупных инфраструктурных проектов информационного обеспечения специалистов в области материаловедения (MGI, MDF, NoMaD и т.д.). Дан краткий обзор российских информационных ресурсов в области неорганической химии и материаловедения. Предложен проект инфраструктуры для обеспечения данными российских специалистов в этой области.

Авторы благодарят А.В. Столяренко, В.В. Рязанова, О.В. Сенько, А.А. Докукина за помощь в создании информационно-аналитической системы.

Работа выполнена при частичной финансовой поддержке РФФИ, проекты 16-07-01028, 14-07-00819 и 15-07-00980.

1 Введение

Конкурентные требования глобального рынка требуют постоянного обновления и улучшения потребительских свойств продукции. Качество и новизна выпускаемой продукции в значительной степени определяется материалами, используемыми при ее производстве. В связи с этим ускорение поиска, исследования и внедрения новых материалов с заданными функциональными свойствами является критически важной задачей развития промышленности и всей экономики стран в целом. В настоящее время, по мнению американских специалистов [1], между открытием нового материала и началом его практического использования проходит более 20 лет. Это связано с тем, что очень часто потребители не имеют достаточной информации даже об очень перспективных материалах, работы по исследованию и созданию технологии получения и обработки материалов необоснованно дублируются, используются не самые лучшие по потребительским и прочим параметрам вещества, что приводит к снижению качества продукции, росту затрат на ее производство и, в конечном счете, к утрате рыночной привлекательности выпускаемого продукта.

Одним из путей ускорения поиска, разработки и внедрения новых материалов является создание развитой инфраструктуры информационного обеспечения специалистов, в первую очередь, распределенной виртуально интегрированной сети баз данных и баз знаний, содержащих информацию о свойствах веществ и материалов и технологиях их получения и обработки, а также систем компьютерного конструирования и моделирования материалов, доступных из Интернет специалистам самого разного профиля: научным работникам, инженерам, технологам, бизнесменам, госслужащим, студентам и т.д.

В последние годы в развитых странах были выдвинуты и поддержаны правительствами инициативы, направленные на организацию инфраструктуры доступа к экспериментальным и расчетным данным о материалах. Краткий обзор некоторых инициатив ранее был дан в [2].

2 Стратегическая Инициатива Геном Материалов (Materials Genome Initiative (MGI))

В 2011 г в США была начата разработка проекта, названного Инициативой Геном Материалов (Materials Genome Initiative (MGI)) [3]. Цели MGI - ускоренное создание новых материалов, обладающих заданными свойствами, что критично для достижения высокого уровня конкурентоспособности промышленности США и будет способствовать поддержке их лидирующей роли во многих секторах современного материаловедения и промышленности: от энергетики до электроники, от обороны до здравоохранения. Особое внимание в MGI уделяется поддержке прорывных исследований в теории, моделировании свойств материалов и data mining как средств достижения существенного прогресса в материаловедении, что приведет к снижению затрат

Труды XVIII Международной конференции DAMDID/RCDL'2016 «Аналитика и управление данными в областях с интенсивным использованием данных», Ершово, 11-14 октября 2016

на разработку, исследование и получение новых материалов. Задачи MGI – обеспечение разработки и внедрения новых материалов, в том числе и за счет координации исследований и предоставления доступа к расчетным моделям и инструментарию для оценки свойств и поведения материалов, а также использования прорывных методов моделирования и анализа данных. Реализация проекта MGI позволит создать механизмы, способствующие обмену данными и знаниями о материалах не только между исследователями, но и между академической наукой и промышленностью. Основой MGI является Инфраструктура инноваций в материаловедении (Materials Innovation Infrastructure), которая обеспечивает интеграцию методов и средств современного моделирования и экспериментальных исследований. Инфраструктура включает комплекс взаимосвязанных обслуживающих структур и объектов (в том числе и установок megascience), составляющих и/или обеспечивающих основу функционирования материаловедения как науки и прикладной области. На первом этапе на реализацию программы MGI выделено около 400 млн. долларов. В Подкомитет MGI Национального научно-технического Совета США (The National Science and Technology Council (NSTC)) входят представители Министерства обороны, Министерства энергетики, National Institute of Standards and Technology (NIST), National Science Foundation (NSF), National Aeronautics and Space Administration (NASA), National Institutes of Health (NIH), United States Geological Survey (USGS), Defense Advanced Research Projects Agency (DARPA) и т.д. [4]. Среди успешных поддержанных проектов этой инициативы можно выделить систему AFLOW [5, 6], содержащую БД с результатами квантовомеханических расчетов веществ и оснащенную компьютерным пакетом программ для проведения таких расчетов, и открытие нового вида высокопрочного и износостойкого стекла [7], осуществленного путем широкого использования теоретических расчетов.

3 Средства организации данных о материалах (The Materials Data Facility (MDF))

Учитывая важность материалов для достижения высокого уровня конкурентоспособности промышленности США, в июне 2014 г. национальный Консорциум сервисов данных (National Data Service (NDS)) объявил о пилотном проекте разработки средств для организации данных о материалах: The Materials Data Facility (MDF) [8], поддерживаемом NIST. Этот проект является ответом на инициативу MGI Белого дома по ускорению разработки современных материалов. MDF обеспечит материаловедов масштабируемым репозиторием для хранения экспериментальных и расчетных данных, в том числе и до их публикации, снабженных ссылками на соответствующие библиографические источники. MDF станет рычагом для создания национальной инфраструктуры коллективного использования информации, включая разработанные в мире БД по свойствам материалов и информационные системы для расчета и моделирования, а также будет способствовать организации обмена данными о материалах, в том числе и еще не опубликованными. Доступность данных и средств расчета обеспечивается современной информационной и телекоммуникационной инфраструктурой, которая позволяет предоставить данные исследователям материалов для многоцелевого использования, дополнительного анализа и проверки. Помимо NIST, среди исполнителей MDF необходимо выделить University of Chicago, Argonne National Laboratory, The University of Illinois, Northwestern University, Center for Hierarchical Materials Design и т.д. Репозиторий MDF сейчас включает [8], помимо многочисленных БД NIST [9], информационные системы с результатами квантовомеханических расчетов: AFLOW [5, 6], The Open Quantum Materials Database (OQMD) [10] и т.д.

4 Программа Поиска Новых Материалов (Novel Materials Discovery Laboratory (NoMaD))

Эта программа был ответом Евросоюза на американскую стратегическую инициативу MGI. Проект NoMaD [11, 12] направлен на создание Европейских центров превосходства (European Centres of Excellence) и предполагает разработку сети БД (Materials Encyclopedia) по свойствам веществ и материалов (в первую очередь, содержащих результаты расчетов), а также средств анализа этих данных и расчета веществ. Цель - ускорение разработки и использования материалов с заданными функциональными свойствами. Программа стартовала в ноябре 2015 г. в рамках проекта ЕС HORIZON2020 (объем финансирования около 5 млн. евро) [12]. Существенным недостатком NoMaD является ориентация на информационные ресурсы США (главным образом, БД NIST по свойствам веществ и материалов) и информационные системы с расчетными данными. В настоящее время репозиторий NoMaD [13] содержит только результаты квантовомеханических расчетов уже полученных соединений. Программа NoMaD во многом коррелирует с проектом Евросоюза Materials design at the eXascale (MaX) [14], включающим создание инфраструктуры для проведения квантовомеханических расчетов с использованием высокопроизводительных компьютерных систем (объем финансирования – свыше 4 млн. евро). Среди исполнителей NoMaD следует отметить ведущие организации Европы, такие как Humboldt University, Fritz-Haber-Institute of the Max Planck Society, King's College London, University of Barcelona, Aaalto University, Max Planck Institute for the Structure of Dynamics of Matter, Technical University of Denmark,

Max Planck Computing and Data Facility, Barcelona Supercomputing Centre и т.д.

5 Инициатива исследования материалов путем интеграции информации («Materials research by Information Integration Initiative» (MI2I))

Эта инициатива была предложена в 2015 г. японским правительством, которое создало на базе Национального института материаловедения (National Institute for Materials Science (NIMS)) Center for Materials Research by Information Integration [15]. В отличие от европейских программ созданный центр ставит своей задачей не только широкое использование квантовомеханических расчетов, но и поддержку и развитие имеющихся в Японии БД по свойствам веществ и материалов [16], их интеграцию с зарубежными информационными системами и применение методов искусственного интеллекта для прогноза новых веществ [17, 18].

6 Анализ реализуемых в мире крупных инфраструктурных проектов информационного обеспечения в области материаловедения

Следует отметить общие тенденции в разработке систем информационного обеспечения в материаловедческих областях:

- создание интегрированной сети БД по свойствам веществ и материалов;
- разработка и широкое применение расчетных методов;
- создание БД с расчетной информацией.

Анализ целей и предлагаемых в вышеуказанных инициативах методов и технологий их достижения показывает, что наиболее перспективны проекты США. Именно они позволят создать полноценную инфраструктуру информационной поддержки инновационной деятельности в разработке и внедрении новых материалов, обеспечив науку и промышленность достоверными и полными данными о свойствах веществ и материалов и разнообразным инструментарием (пакеты квантовомеханических расчетов, data mining и т.д.) для расчетов параметров веществ. Японская инициатива более ограничена. Она основана на использовании системы БД по свойствам веществ и материалов NIMS, а также использует имеющийся у исполнителей задел по применению уже известных расчетных методов (например, широко известного пакета квантовомеханических расчетов VASP [19]). Начаты работы по применению методов искусственного интеллекта [17, 18]. К тому же японские специалисты ограничили сферу деятельности материалами для электроники (источники питания, магнитные, термоэлектрические и спинтронные материалы) [15]. Проекты ЕС на их начальном этапе выглядят наименее перспективными. Ориентация на американские БД по свойствам веществ и только квантовомеханические расчеты значительно снижают потенциал и возможности этих инфраструктурных проектов. Тем не менее, следует отметить, что объективной предпосылкой для успешной реализации предложенных в США, ЕС и Японии инициатив являются, с одной стороны, успехи в разработке и применении методов расчета свойств веществ, и, с другой стороны, наличие множества баз данных по свойствам веществ и материалов, разработанных в последние годы в разных странах (обзор имеющихся БД в области неорганический химии и материаловедения дан в статье [20] и в БД IRIC (Information Resources on Inorganic Chemistry) [21]). Несмотря на то, что на создание и поддержку таких информационных систем затрачены сотни миллионов долларов, их использование экономически выгодно, т.к. они позволяют значительно сократить затраты на разработку новых материалов за счет уменьшения дублирования исследований и предоставления химикам и материаловедам оперативной и достоверной информации о свойствах веществ. В свою очередь, расчетные методы дают возможность еще до экспериментов оценить параметры веществ, указать перспективные для применений составы и разработать технологию получения и обработки материалов. Следствием решения этих задач является резкое сокращение затрат и времени на разработку и внедрение новых материалов. К сожалению, из-за введенных санкций против РФ, предполагаемой высокой цены доступа к разрабатываемым информационным системам, наличия в них информации о материалах и технологиях двойного назначения, доступ российских специалистов к этим информационным ресурсам будет крайне ограничен. Это может привести к серьезному отставанию в темпах разработки и внедрения новых материалов, что приведет к резкому уменьшению конкурентоспособности российской продукции, особенно в наукоемких отраслях. Единственный путь решения этой проблемы – это создание собственной инфраструктуры, обеспечивающей науку, образование, промышленность, бизнес, административные органы данными о материалах, технологиях их получения и обработки, сферах применения, производителях и потребителях материалов, а также средствами обработки и анализа накопленной информации, компьютерного моделирования и конструирования новых веществ и материалов, позволяющими принимать решения о выборе материалов для конкретных применений, перспективности разработки и использования конкретного вещества, о технологических особенностях производства, использования, утилизации и т.д. материалов и т.п. Средства, вложенные в такую импортозамещающую программу, достаточно быстро окупятся за счет сокращения затрат на разработку и исследование

новых материалов и прибыли от реализации наукоемкой продукции, конкурентной на мировом рынке.

7 Опыт разработки интегрированной информационной системы по свойствам неорганических веществ и материалов

Предпосылкой для успешного выполнения такого инфраструктурного проекта в России является опыт в разработке и интеграции БД по свойствам неорганических веществ и материалов, доступных из сети Интернет, также методов и программных средств для компьютерного конструирования новых веществ и материалов, основанных на использовании технологий data mining, и, в первую очередь, методов распознавания образов по прецедентам [20, 23]. Следует отметить, что интерес к применению методов data mining в неорганическом материаловедении связан с объективными трудностями, возникающими при квантовомеханических расчетах еще не полученных многокомпонентных неорганических веществ, особенно в твердой фазе. Например, для того, чтобы рассчитать электронную структуру неорганического соединения с использованием пакета VASP [19], необходимо знать его кристаллическую структуру, т.е. нужно получить и исследовать это соединение.

Рисунок 1 Схема информационно-аналитической системы для конструирования неорганических соединений

С помощью же методов распознавания образов, проанализировав имеющуюся информацию об уже известных веществах, хранящуюся в БД, можно прогнозировать еще не синтезированные соединения и оценивать некоторые их свойства, зная только хорошо известные параметры компонентов (химических элементов или более простых соединений). Для решения этой задачи в ИМЕТ РАН разработана специальная информационно-аналитическая система (ИАС) (рис.1), включающая интегрированную систему БД по свойствам неорганических веществ и материалов, подсистему поиска закономерностей в данных, прогнозирования новых соединений и оценки их свойств, базу знаний, базу прогнозов и другие подсистемы [22].

7.1 Интегрированная система баз данных по свойствам неорганических веществ и материалов

Интегрированная система баз данных по свойствам неорганических веществ и материалов в настоящее время объединяет информационные системы, разработанные в ИМЕТ РАН [20]: по фазовым диаграммам полупроводниковых систем («Диаграмма»), по свойствам акустооптических, электрооптических и нелинейнооптических веществ («Кристалл»), по ширине запрещенной зоны неорганических веществ («Bandgap»), по свойствам неорганических соединений («Фазы») и по свойствам химических элементов («Elements»), а также БД «AtomWork» по свойствам неорганических веществ, разработанную в National Institute for Materials Science (Япония), и БД ТКВ по термическим константам веществ, разработанную в ОИВТ РАН и МГУ.

БД по свойствам неорганических соединений «Фазы» [25, 26] в настоящее время содержит информацию о свойствах более 52 тыс. тройных соединений (т.е. соединений, образованных тремя химическими элементами) и более 31 тыс. четверных соединений, почерпнутую из более 32 тыс. литературных источников. Она включает краткую информацию о наиболее распространенных свойствах неорганических соединений: кристаллохимических (тип кристаллической структуры с указанием температуры и давления, выше которых реализуется указанная структура, сингония, пространственная группа, число формульных единиц в элементарной ячейке, параметры кристаллической решетки) и теплофизических (тип и температура плавления, температура распада соединения в твердой или газообразной фазах и температура кипения при атмосферном давлении). Помимо этого, БД содержит информацию о сверхпроводящих свойствах соединений. БД «Фазы» формируется на основе анализа сведений, почерпнутых из периодических изданий, справочников, монографий, отчетов, а также реферативных журналов (более половины источников хранятся в виде pdf-документов). Объем БД «Фазы» превышает 25 Гбайт, и она доступна

зарегистрированным пользователям из сети Интернет [25].

БД «Elements» [20, 26] включает информацию о 90 наиболее распространенных свойствах химических элементов: теплофизических (температура плавления и кипения при 1 атм, стандартные теплопроводность, молярная теплоемкость, энтальпия атомизации, энтропия и т.д.), размерных (ионные, ковалентные, металлические, псевдопотенциальные радиусы, объем атома и т.д.), других физических свойствах (магнитной восприимчивости, электропроводности, твердости, плотности и т.д.), положении в Периодической таблице элементов и т.д. БД доступна из сети Интернет [26].

БД «Диаграмма» [27, 28] содержит информацию, собранную и оцененную высококвалифицированными экспертами, о фазовых P-T-x-диаграммах двух- и трехкомпонентных полупроводниковых систем и о физико-химических свойствах образующихся в них фаз. Объем БД превышает 2 Гбайт. БД доступна зарегистрированным пользователям из сети Интернет [27].

БД «Bandgap» [29, 30] включает информацию (более 0.7 Гбайт) о ширине запрещенной зоны более 3 тыс. неорганических веществ. БД доступна пользователям из сети Интернет [30]. По предложению японской стороны эта БД будет интегрирована с японской БД «Computational Electronic Structure Database (CompES-X)», содержащей информацию об электронной структуре веществ [31].

БД «Кристалл» [32, 33] включает информацию о свойствах (пьезоэлектрических (пьезоэлектрические коэффициенты, упругие постоянные и т.д.), нелинейно-оптических (нелинейно-оптические коэффициенты, компоненты тензора Миллера и т.д.), кристаллохимических (тип кристаллической структуры, сингония, пространственная и точечная группа, число формульных единиц в элементарной ячейке, параметры кристаллической решетки), оптических (показатели преломления, область прозрачности и т.д.), теплофизических (температура плавления, теплоемкость, теплопроводность и т.д.) и т.д.), более 140 акустооптических, электрооптических и нелинейно-оптических веществ, собранную и оцененную высококвалифицированными экспертами в данной предметной области. Объем БД превышает 4 Гбайт. Она имеет русско- и англоязычную версии, доступные зарегистрированным пользователям из сети Интернет [32].

БД «Inorganic Material Database – AtomWork» [34, 35] содержит информацию о более чем 82 тыс. кристаллических структур, 55 тыс. значений свойств материалов и 15 тыс. фазовых диаграмм. БД доступна пользователям из сети Интернет [35].

БД по термическим константам веществ «ТКВ» [36] содержит доступную из сети Интернет информацию об около 27 тыс. веществ, образованных практически всеми химическими элементами.

7.2 Система компьютерного конструирования неорганических соединений

Основу системы компьютерного конструирования неорганических соединений составляют алгоритмы и программы распознавания образов по прецедентам, входящие в многофункциональную систему «РАСПОЗНАВАНИЕ», разработанную в ВЦ РАН [39] и объединяющую, помимо широко известных методов линейной машины, линейного дискриминанта Фишера, k-ближайших соседей, опорных векторов, нейросетевых и генетических алгоритмов, также уникальные алгоритмы, разработанные в ВЦ РАН: алгоритмы распознавания, основанные на вычисления оценок, алгоритмы голосования по тупиковым тестам, алгоритмы голосования по логическим закономерностям, алгоритмы статистического взвешенного голосования и т.д. В систему также интегрирована программа обучения ЭВМ процессу формирования понятий ConFor, разработанная в Институте кибернетики НАН Украины [38], в основу которой положена оригинальная организация данных в памяти ЭВМ в виде растущих пирамидальных сетей. Для отбора информативных свойств компонентов химических соединений в ИАС были включены программы, основанные на алгоритмах [39-41]. Использование методов распознавания образов позволило получить прогнозы тысяч новых неорганических соединений [22, 23, 29].

8 Проект инфраструктуры обеспечения данными российских специалистов в области неорганической химии и материаловедения

ИАС является, своего рода, пилотным проектом для создания информационной инфраструктуры для неорганического материаловедения. В ней виртуально интегрированы наиболее известные российские БД в этой области, а также начата их интеграция с зарубежными информационными системами. Большинство российских БД содержат ссылки на полные тексты публикаций, из которых извлечена информация, хранящаяся в БД. Подсистема компьютерного конструирования соединений позволяет найти закономерности в информации БД и использовать их для прогнозирования еще не полученных соединений и оценки их свойств. Следует отметить, что на этапе прогнозирования используются только данные о свойствах компонентов соединений (химических элементов или более простых соединений). Полученные прогнозы хранятся в специальной базе прогнозов, что расширяет функциональные возможности традиционных баз данных (пользователь получает не только известные

экспериментальные данные, но и прогнозы еще не синтезированных соединений и оценки некоторых их свойств).

При разработке российского проекта инфраструктуры информационного обеспечения специалистов в области неорганического материаловедения нужно учитывать все многообразие возможных запросов пользователей. Вполне естественно, что запросы академических ученых могут кардинально отличаться от запросов инженеров-конструкторов или производителей материалов. Однако общий проект информационной инфраструктуры должен включать в качестве необходимых элементов виртуально интегрированную систему российских и зарубежных баз данных по свойствам неорганических веществ и материалов, технологиям их получения и обработки, потребителям и производителям материалов и т.д., комплекс пакетов расчета и моделирования материалов, пользователями которых в большинстве случаев являются академические ученые, и виртуально интегрированную систему баз данных уже рассчитанных значений (рис. 2). Следует подчеркнуть, что технологии обработки, хранения, организации поиска необходимых сведений требуют разработки и использования самых современных программных средств и создания мощных центров обработки данных (ЦОД).

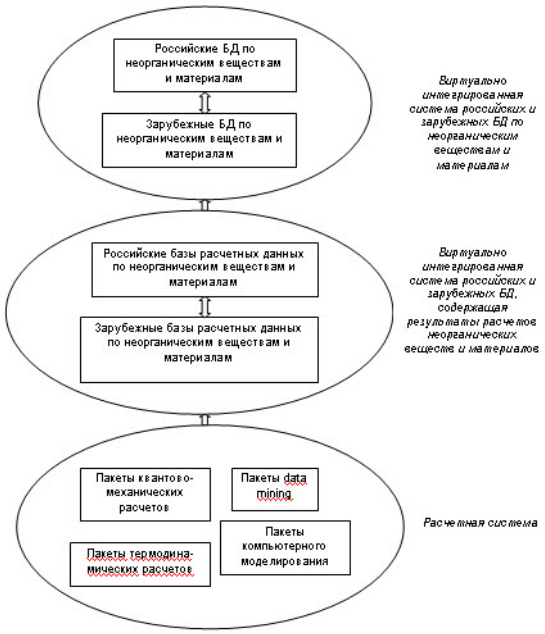


Рисунок 2 Схема инфраструктуры обеспечения данными российских специалистов в области неорганической химии и материаловедения

Система БД должна виртуально интегрировать наиболее важные для российских пользователей фактографические БД по неорганическим веществам и материалам (российские БД ИМЕТ РАН, ОИВТ РАН, МГУ и т.д и зарубежные БД NIMS [16], NIST [9], STN [42], Springer Materials [43], Materials Science International (MSI) [44] и т.д.), документальные БД ведущих издательских корпораций (Наука, Elsevier, Springer, Wiley, American Chemical Society, American Institute of Physics, Science и т.д.), а также базы еще не опубликованных в открытой печати сообщений (ВИНИТИ, ЦИТИС и т.д.), патентные базы (Роспатент, Questel, USPTO и т.д.), базы потребителей и производителей неорганических материалов и т.д. Необходимо выделять средства на ежегодное продление лицензий для пользования зарубежными базами и организовать единый портал бесплатного для российских пользователей доступа к ним (сейчас такие базы доступны для ограниченного числа организаций). Необходимо всячески поддерживать перевод в электронную форму коллекций наиболее известных в мире российских журналов, что несомненно будет способствовать повышению их авторитета и цитируемости.

Оснащение исследовательских организаций системами расчета необходимо начинать, в первую очередь, с обучения студентов и аспирантов пользованию наиболее известными пакетами квантовомеханических, термодинамических, статистических и т.д. расчетов. К сожалению, за последние четверть века отток специалистов в области теоретической химии сильно обескровил всемирно известные школы российских университетов и академических институтов, что значительно усложняет решение проблемы квалифицированного обучения молодых специалистов, а также использования расчетных методов. Нужно разрабатывать российские базы данных расчетных значений и интегрировать их с зарубежными информационными системами, которые сейчас еще открыто доступны в Интернет (например, [13, 30]), что позволит частично решить проблему квалифицированных расчетов веществ. Постановка экспериментов должна включать проведение или использование расчетов в качестве начального этапа исследований, что позволит сократить время и затраты на поиск и разработку новых материалов.

Важным компонентом разрабатываемого инфраструктурного проекта должна стать подсистема анализа запросов пользователей, особенно, специалистов в прикладных областях. Именно она позволит выявить группы материалов, на изучение которых нужно направить экспериментальные исследования. Статистика отказов в выдаче значений того или иного параметра вещества может стать стимулом к дополнительному экспериментальному исследованию запрашиваемого свойства.

9 Заключение

Переход российской экономики на инновационный путь развития и повышение конкурентоспособности продукции во многом определяется качеством, новизной и функциональными возможностями материалов. На

современном этапе развития технологий поиск, исследование и внедрение новых материалов требует создания развитой инфраструктуры, включающей академические организации с их потенциалом теоретических и экспериментальных исследований новых веществ, организации, ведущие прикладные исследования по разработке и внедрению новых материалов и технологий их получения и обработки, центры коллективного пользования с комплексами дорогостоящих установок, включая объекты megascience, и т.д. В последние годы в развитых странах инициированы стратегически важные для обеспечения технологического превосходства проекты (MGI, MDF, NoMaD и т.п.) создания инфраструктуры для ускорения разработки и внедрения новых материалов, обладающих заданными функциональными свойствами. Особое внимание в этих проектах уделено инфраструктуре информационного обеспечения. Российским ответом на стратегические инициативы США, ЕС, Японии в плане информационной инфраструктуры может являться создание федерального информационного центра, обеспечивающего специалистов информацией о свойствах веществ и материалов, технологиях их производства, а также расчетными данными, патентной информацией и т.д. В связи со спецификой предметной области основу такого информационного центра коллективного пользования должна составлять распределенная виртуально интегрированная сеть отечественных и зарубежных баз данных и баз знаний по веществам и материалам. Создание федерального информационного центра, интегрирующего информационные ресурсы в области материаловедения, будет способствовать резкому ускорению поиска, разработки и внедрения новых материалов, в сочетании со значительным сокращением затрат за счет уменьшения дублирования исследований и предоставления химикам и материаловедам оперативной и достоверной экспериментальной и расчетной информации о веществах и материалах.

Литература

[1] Materials Genome Initiative. Strategic Plan. National Science and Technology Council. Committee on Technology. Subcommittee on the Materials Genome Initiative. December 2014. 55 p. https://www.whitehouse.gov/sites/default/files/microsites/ostp/NSTC/mgi_strategic_plan_-_dec_2014.pdf

[2] Калиниченко Л.А., Вольнова А.А., Гордов Е.П., Киселева Н.Н. и др., Проблемы доступа к данным в исследованиях с интенсивным использованием данных в России. Информатика и ее применения, 10(1), с. 3-23, 2016.

[3] Materials Genome Initiative for Global Competitiveness/ June 2011. 18 p. http://www.whitehouse.gov/sites/default/files/microsites/ostp/materials_genome_initiative-final.pdf

[4] Materials Genome Initiative. https://www.mgi.gov/partners

[5] Curtarolo S., Setyawan W., Wang S., et al. AFLOWLIB.ORG: A distributed materials properties repository from high-throughput ab initio calculations. Comp. Mat. Sci., 58, p.227-235, 2012.

[6] Taylor R. H., Rose F., Toher C., et al. RESTful API for exchanging materials data in the AFLOWLIB.org consortium. Comp. Mat. Sci., 93, p. 178-192, 2014.

[7] Сайт University of Chicago http://www.uchicago.edu/features/microscopic_animals_inspire_innovative_glass_research/

[8] National Data Service. The Materials Data Facility. http://www.nationaldataservice.org/mdf/

[9] NIST Data Gateway. http://srdata.nist.gov/gateway/gateway?dblist=0

[10] Saal, J. E., Kirklin, S., Aykol, M., et al. Materials Design and Discovery with High-Throughput Density Functional Theory: The Open Quantum Materials Database (OQMD), JOM. 65, p. 1501–1509, 2013.

[11] The Novel Materials Discovery (NOMAD) Laboratory. http://nomad-lab.eu/

[12] The Novel Materials Discovery (NOMAD) Laboratory. http://cordis.europa.eu/project/rcn/198339_en.html

[13] The NoMaD Repository. http://nomad-repository.eu/cms/

[14] Materials design at the eXascale. http://cordis.europa.eu/project/rcn/198340_en.html/

[15] Center for Materials Research by Information Integration. http://www.nims.go.jp/eng/research/MII-I/index.html

[16] NIMS Materials Database (MatNavi). http://mits.nims.go.jp/index_en.html

[17] Lee J., Seko A., Shitara K., Tanaka I. Prediction model of band-gap for AX binary compounds by combination of density functional theory calculations and machine learning techniques. Phys. Rev., B93(11), p. 115104, 2016.

[18] Toyoura K., Hirano D., Seko A., et al. Machine-learning-based selective sampling procedure for identifying the low-energy region in a potential energy surface: A case study on proton conduction in oxides. Phys. Rev., B93(5), p. 054112, 2016.

[19] VASP-site. https://www.vasp.at/

[20] Киселева Н.Н., Дударев В. А., Земсков В. С. Компьютерные информационные ресурсы неорганической химии и материаловедения. Успехи химии. 79(2), с. 162-188, 2010.

[21] БД IRIC (Information Resources on Inorganic Chemistry). http://iric.imet-db.ru/

[22] Киселева Н.Н. Компьютерное конструирование неорганических соединений. Использование баз данных и методов искусственного интеллекта. М.: Наука, 2005.

[23] Киселева Н.Н., Дударев В.А., Столяренко А.В. Интегрированная система баз данных по свойствам неорганических веществ и материалов. Теплофизика высоких температур, 54(2), с. 228–236, 2016.

[24] Киселева Н., Мурат Д., Столяренко А. и др. База данных по свойствам тройных неорганических соединений «Фазы» в сети Интернет. Информационные ресурсы России, 4, с. 21-23, 2006.

[25] БД «Фазы». www.phases.imet-db.ru

[26] БД «Elements». http://phases.imet-db.ru/elements

[27] 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, с. 50-55, 2001.

[28] БД "Диаграмма". http://diag.imet-db.ru

[29] Киселева Н.Н., Дударев В.А., Коржуев М.А. База данных по ширине запрещенной зоны неорганических веществ и материалов. Материаловедение, 7, с. 3-8, 2015.

[30] БД «Bandgap». http://www.bg.imet-db.ru

[31] DB CompES-X. http://compes-x.nims.go.jp/index_en.html

[32] 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), с.380-384, 2004.

[33] БД "Кристалл". http://crystal.imet-db.ru

[34] Xu Y., Yamazaki M., Villars P. Inorganic Materials Database for Exploring the Nature of Material. Jap. J. Appl. Phys., 50(11), p.11RH02-1-5, 2011.

[35] БД "AtomWork". http://crystdb.nims.go.jp/index_en.html

[36] БД "ТКВ". http://www.chem.msu.su/cgi-bin/tkv.pl?show=welcome.html/welcome.html

[37] Журавлев Ю. И., Рязанов В. В., Сенько О. В. «РАСПОЗНАВАНИЕ». Математические методы. Программная система. Практические применения. М.: ФАЗИС, 2006.

[38] Гладун В. П. Процессы формирования новых знаний. София: СД "Педагог-6", 1995.

[39] Senko O. V. An Optimal Ensemble of Predictors in Convex Correcting Procedures // Pattern Recognition and Image Analysis, 19(3), p. 465-468, 2009.

[40] Yuan G.-X., Ho C.-H., Lin C.-J. An Improved GLMNET for L1-regularized Logistic Regression // J. Machine Learning Research, 13, p. 1999-2030, 2012.

[41] Yang Y., Zou H. A Coordinate Majorization Descent Algorithm for L1 Penalized Learning // J. Statistical Computation&Simulation, 84(1), p. 1-12, 2014.

[42] Сайт STN. http://www.stn-international.de/stn_home.html?&L=snavidlizydkyfaz

[43] Сайт Springer Materials. http://materials.springer.com/welcome

[44] Сайт MSI. http://www.msiport.com/.

Inorganic Chemistry and Materials Science Data Infrastructure for Specialists

Nadezhda N. Kiselyova, Victor A. Dudarev

World-wide materials science infrastructure projects are analyzed (MGI, MDF, NoMaD, etc.). Short overview of the Russian information resources on inorganic chemistry and materials science is given. Infrastructure project is proposed for Russian specialists in the domain to provide them with data.

Интеграция пользовательских интерфейсов информационных систем в области неорганической химии

© В.А. Дударев[1,2] © Н.Н. Киселева[1]
[1]ФГБУН Институт металлургии и материаловедения им. А.А. Байкова РАН (ИМЕТ РАН),
[2]Национальный исследовательский университет «Высшая школа экономики» (НИУ ВШЭ), Москва
vic@imet.ac.ru kis@imet.ac.ru

Аннотация

Часто при поиске данных по свойству того или иного вещества неискушенный пользователь не знает к какой информационной системе по свойствам неорганических веществ и материалов (ИС СНВМ) стоит прибегнуть для первичного сбора информации. Поэтому актуальным является создание специализированной ИС, позволяющей потребителю данных по свойствам неорганических веществ получить возможность просмотра связанной информации по свойствам заданной химической системы в разных ИС СНВМ из одного места, которое условно называется “единой точкой входа”. Созданию именно такой ИС, являющейся единой точкой входа для пользователя в ИС СНВМ, посвящена настоящая работа [1].

Работа выполнена при частичной финансовой поддержке РФФИ, проекты 16-07-01028, 14-07-00819 и 15-07-00980.

1 Поиск релевантной информации

Основная идея заключается в предоставлении пользователю возможности выбора химических элементов (из периодической таблицы Менделеева), образующих химическую систему. Имея набор выбранных пользователем элементов, ИС единой точки входа должна осуществить поиск ИС СНВМ, содержащих сведения о свойствах фаз выбранной химической системы, для чего используется метабаза – специализированная база данных, хранящая справочные сведения о содержимом интегрируемых ИС СНВМ. Метабаза и механизмы ее наполнения были разработаны ранее при создании интегрированной ИС СНВМ [2].

Опишем содержимое метабазы в терминах теории множеств [3]. В метабазе содержится информация по интегрируемым ИС (множество D), химическим системам (множество S) и их свойствам (множество P). Для описания взаимосвязи между элементами множеств D, S и P было определено тернарное отношение W на множестве $U = D \times S \times P$. Принадлежность элемента (d, s, p) отношению W, где $d \in D, s \in S, p \in P$, интерпретируется следующим образом: “в интегрируемой ИС d содержится информация по свойству p химической системы s”.

Поиск релевантной информации s сводится к определению вида отношения R, являющегося подмножеством декартова произведения $S \times S$ (иными словами, $R \subset S^2$). Таким образом, о любой паре $(s_1, s_2) \in R$ можно сказать, что система s_2 является релевантной системе s_1. Т.е., чтобы решить задачу поиска релевантной информации в интегрируемых информационных системах, необходимо определить отношение R.

При построении ИС единой точки входа в ИС СНВМ отношение релевантности строится следующим образом: для любых множеств $s_1 \in S, s_2 \in S$, состоящих из химических элементов e_{ij}, $s_1 = \{e_{11}, e_{12}, .., e_{1n}\}, s_2 = \{e_{21}, e_{22}, .., e_{2m}\}$ верно, что если $s_1 = s_2$, то $(s_1, s_2) \in R$. Как видно из условия, отношение R симметрично. Таким образом, получим в качестве релевантных только те химические системы, вещества и модификации, которые состоят из одного и того же набора химических элементов (одинаковые химические системы). Как правило, этот способ построения отношения R является наиболее часто используемым при поиске всех свойств заданного химического вещества или системы через единую точку входа.

Поскольку поиск релевантной информации выполняется в метабазе, единая точка входа

Труды XVIII Международной конференции DAMDID/RCDL’2016 «Аналитика и управление данными в областях с интенсивным использованием данных», Ершово, 11-14 октября 2016

предоставляется для всех ИС СНВМ, описанных в метабазе. В настоящее время интегрированная ИС СНВМ консолидирует все разработанные в ИМЕТ РАН информационные системы: Фазы, Elements, Диаграмма, Кристалл и Bandgap [4], а также ИС “ТКВ” по термическим константам веществ, разработанную в ОИВТ РАН и МГУ. Благодаря проделанной работе по международной интеграции, удалось включить в состав интегрированной системы ИМЕТ РАН ИС AtomWork (разработанную в National Institute for Materials Science (NIMS), Япония), содержащую информацию о более чем 23 тыс. неорганических веществ [5, 6]. Потенциально, в состав интегрированной ИС СНВМ можно включить все ИС из списка наиболее значимых ИС в области неорганической химии и материаловедения [7].

2 Разработка Web-приложения ИС

Рассмотрим кратко особенности разработки Web-приложения единой точки входа, располагаемого по адресу http://meta.imet-db.ru. Web-приложение ASP.Net написано на языке C# (.Net Framework 4.5) с использованием ADO.Net для доступа к метабазе. Для построения запросов используются языковые средства Transact-SQL, являющегося диалектом языка Structured Query Language (SQL), который используется в системе управления базами данных (СУБД) Microsoft SQL Server 2014.

Пользовательский интерфейс является интерактивным за счет использования библиотеки jQuery, облегчающей взаимодействие с HTML DOM (Document Object Model – объектная модель документа) и предоставляющей удобный интерфейс (API) для работы с AJAX (Asynchronous Javascript and XML – асинхронный JavaScript и XML).

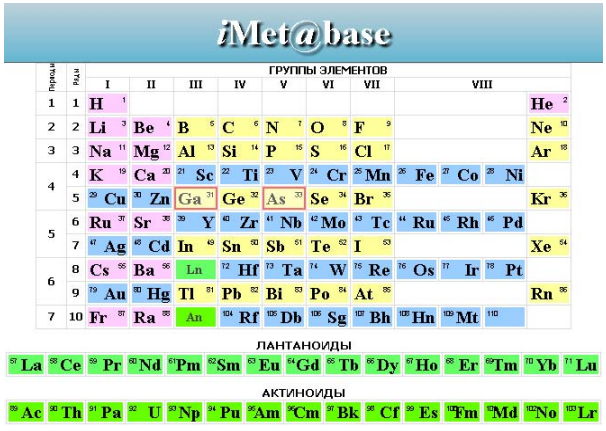


Рисунок 1 Выбор химической системы по набору элементов

Опишем принцип работы Web-приложения. Основной элемент интерфейса пользователя – интерактивная таблица Менделеева (рис. 1). Пользователю предоставляется возможность выбора химических элементов из таблицы Менделеева, образующих химическую систему. При нажатии на каждый химический элемент (выбор или снятие выбора) происходит его подсветка за счет применения классов из каскадных таблиц стилей (CSS) с помощью библиотеки jQuery (язык программирования JavaScript):

```
$(".Mendeleev .element").click(function()
{ // клик на элементе
    $(this).toggleClass("selected"); //
переключаем класс
    var arr = [];

$(".Mendeleev .selected").each( function(
)
{ arr.push($(this).children(".name").text
());
    });
    arr.sort();
    var st = arr.join("-");
    $(".result").html("");
    if (st == "") {
        $(".Mendeleev .inactive").show();
        $(".Mendeleev .active").hide();
    }
    else {
        $(".Mendeleev .inactive").hide();
        $(".Mendeleev .active").show();
        ProcElements(st);
    }

$(".Mendeleev .selectedSystem").html(st);
});
```

Одновременно в случае наличия выбранных химических элементов происходит вызов функции `ProcElements`, которая обеспечивает отправку асинхронного AJAX-запроса к HTTP-обработчику http://meta.imet-db.ru/JSON_Elements.ashx, являющемуся сервисом поиска релевантной информации:

```
function ProcElements(elements) {
    $(".result").html("<center><img
src='/i/loaderlight.gif'
alt='подождите...' width=24 height=24
/></center>");
    $.ajax({
        type: "post",
        url: "/JSON_Elements.ashx",
        data: { "mode":
"getelementsinfo", "elements":
elements },
        dataType: "json",
        error: function(XMLHttpRequest,
textStatus, errorThrown) {

$(".result").html("<center><span
class='err'>ajax error textStatus=" +
textStatus + ", errorThrown=" +
errorThrown + "</span></center>");
        },
        success: function(json) {
            if (json.MsgRu != "" ||
json.MsgEn != "") {

$(".result").html("<center><span
class='err ru'>" + json.MsgRu +
"</span><span class='err en'>" +
json.MsgEn + "</span></center>")
                return;
            }
```

```
            var st = "";
            if
(json.Data.Table[0].Row.length == 0) {
                st = "<center><span
class='err ru'>нет данных</span><span
class='err en'>нет
данных</span></center>";
            }
            else {
                st =
RenderDataAsTable_PopUp(json); // список-
попап
            }
            $(".result").html(st);
        }
    });
}
```

При выборе, например, химической системы As-Ga будет отправлен следующий POST запрос на адрес: http://meta.imet-db.ru/JSON_Elements.ashx содержащий данные mode=getelementsinfo&elements=As-Ga. Задача сервиса поиска релевантной информации – по множеству выбранных химических элементов вернуть перечень ИС СНВМ, содержащих сведения о заданной химический системе. Поскольку клиентская часть отработки информации реализована на JavaScript, как стандартном языке сценариев, поддерживаемом всеми браузерами, то сервис поиска формирует ответ в формате JSON (JavaScript Object Notation – нотация объектов JavaScript), который является естественным при использовании языка JavaScript. Приведем пример возвращаемого документа при поиске информации по элементу Ga (исключительно для краткости, т.к. по As-Ga размер JSON документа на порядок больше):

```
{"MsgRu":"","MsgEn":"","Data":{ "Table":[
{ "Row":[ { "Col":[ "5","31","-Ga-
","1","Ga","","1","Свойства
элемента","","/elements/properties_all_gi
ven.aspx?elem=#IDS#","Элементы","http://p
hases.imet-
db.ru/elements/","","post","http://phases
.imet-
db.ru/elements/GateElements.asp?chk=#CHKS
UM#&t=#TOKEN#","ru"]},
{ "Col":[ "7","16027","-Ga-
","1","Ga","","1","Information","","","At
omWork (NIMS,
Japan)","http://crystdb.nims.go.jp/index_
en.html","","link","https://login-
matnavi.nims.go.jp/sso/UI/Login?goto=http
%3A%2F%2Fcrystdb.nims.go.jp%2Fcrystdb%2Fs
earch-materials-
list%3FisVisiblePeriodicTable%3Dtrue%26co
ndition_type%3Dchemical_system%26need_mor
e_type%3Dprototype_number%26condition_val
ue%3D#ELEMENTS_PLUSinURL#&IDToken1=#CHKSU
M#&IDToken1=#TOKEN#","en"]} ]}]}}
```

Как видно, полученный JSON-документ возвращает в числе прочих данных ссылки для перехода на ИС СНВМ с релевантной информацией. Однако ссылки нуждаются в дальнейшей обработке в частности для замены “#CHKSUM#” и “#TOKEN#”, которые являются контрольной суммой (вычисленной с использованием хеш-функции MD5) с отпечатком параметров перехода и маркером безопасности (token) соответственно. За обработку и вывод пользователю информации из JSON-документа отвечает функция `RenderDataAsTable_PopUp(json)`, результат работы которой виден на рис. 2.

Рисунок 2 Список релевантной информации в ИС СНВМ для системы As-Ga

При наведении пользователем указателя на химическую сущность (систему, вещество или кристаллическую модификацию) выводится список свойств, доступных для просмотра в соответствующей ИС СНВМ (рис. 3).

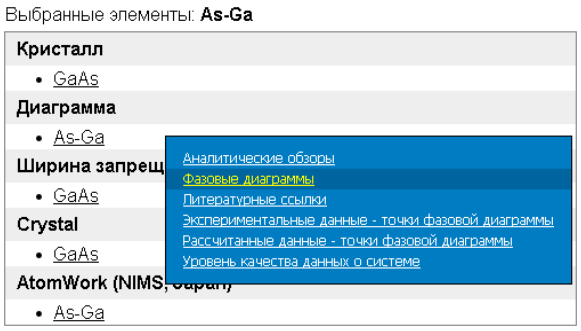


Рисунок 3 Список свойств в ИС “Диаграмма” для системы As-Ga.

Пользователь может, щелкнув по гиперссылке, прозрачно перейти через шлюз безопасности единой точки входа http://meta.imet-db.ru/gate/gateSAP.aspx, в ИС СНВМ с запрошенной информацией. При этом происходит автоматическое перенаправление на страницу с требуемой информацией, например, при переходе по ссылке “Фазовые диаграммы” пользователь увидит сразу затребованную информацию (рис. 4).

Таким образом, в работе на основе метода интеграции информационных систем (Enterprise Application Integration, EAI) реализована единая точка входа во все ИС СНВМ, описанные в каталоге информационных ресурсов метабазы (рис. 5). Желтыми стрелками на рисунке показаны потоки данных при запросах релевантной информации, зелеными стрелками показан переход пользователя

из единой точки входа в ИС СНВМ с релевантной информацией, а синими стрелками обозначен переход пользователя из контекста одной из ИС СНВМ в контекст ИС СНВМ с релевантной информацией.

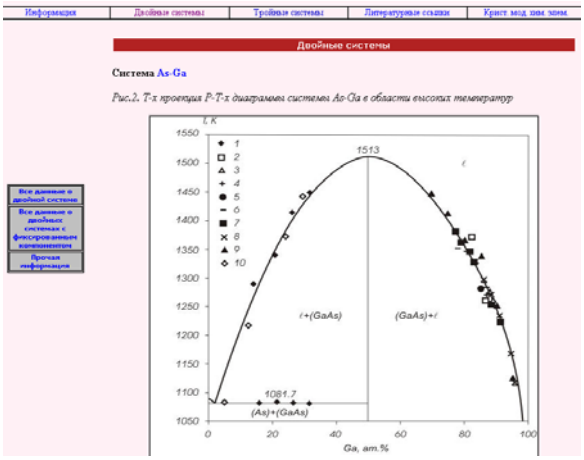


Рисунок 4 Фазовая диаграмма для системы As-Ga в ИС “Диаграмма”

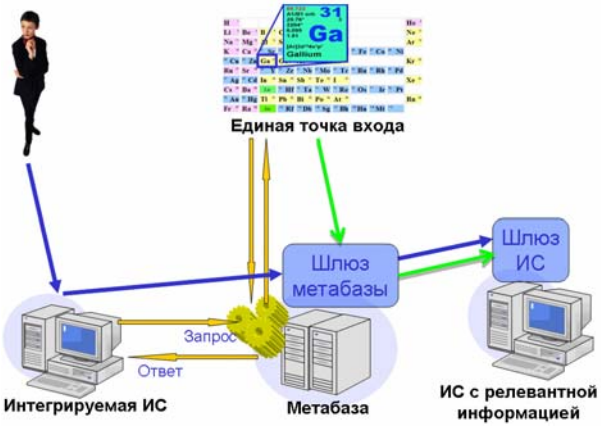


Рисунок 5 Реализации интеграции ИС СНВМ методом EAI

3 Заключение

На текущий момент разработанная интегрированная ИС СНВМ является единственной успешной попыткой интеграции материаловедческих ИС на территории России. Достоверность приведенных в работе выводов подтверждается практической реализацией интегрированной ИС, которая может использоваться конечными пользователями для поиска и сбора информации (EAI) по свойствам неорганических веществ и материалов.

Литература

[1] Дударев В.А. Единая точка входа в информационные системы по свойствам неорганических веществ // X Российская ежегодная конференция молодых научных сотрудников и аспирантов “Физико-химия и технология неорганических материалов”. Сборник материалов. М.: ИМЕТ РАН, 2013. С. 84-86.

[2] Дударев В.А., Филоретова О.А. Подход к интеграции баз данных по свойствам неорганических веществ на основе метабазы // Прикладная информатика, №4(46), 2013, с. 38–42.

[3] Kornyshko V., Dudarev V. Software Development for Distributed System of Russian Databases on Electronics Materials // Int. Journal “Information Theories & Applications”, vol. 13, number 2, 2006. P. 121-126.

[4] Киселева Н.Н., Дударев В.А. Информационная система по ресурсам неорганической химии и материаловедения // Вестник Казанского технологического университета, Т. 17, №19, 2014, с. 356-358.

[5] Дударев В.А. Международная интеграция баз данных по свойствам неорганических веществ // VIII Российская ежегодная конференция молодых научных сотрудников и аспирантов “Физико-химия и технология неорганических материалов”. Сборник материалов. М.: ИМЕТ РАН, 2011. стр: 158-159.

[6] Dudarev V.A., Kiselyova N.N., Xu Y., Yamazaki M. Virtual integration of the Russian and Japanese databases on properties of inorganic substances and materials // MITS 2009. Symposium on Materials Database. National Institute for Materials Science (NIMS). Materials Database Station (MDBS). 2009. p. 37-48.

[7] Киселева Н.Н., Дударев В.А. База данных "Информационные ресурсы неорганической химии и материаловедения" // Информационные технологии, 2010, № 12, с. 63-66.

User interface integration for the information systems on inorganic chemistry

Victor A. Dudarev, Nadezhda N. Kiselyova

Often unexperienced user faces difficulties on materials science information search since it is unknown what information system should be used. Therefore, development of single access point application is of great importance for users searching for materials science information since it allows them to browse relevant data in heterogeneous information systems.

Приглашенный доклад

Invited Talk

Text Mining bridging the gap between knowledge and text (Extended Abstract)

Sophia Ananiadou
NaCTeM, University of Manchecter
kis@imet.ac.ru

Useful pathway models require a complete and accurate representation of the system, which requires that all relevant molecular species are captured, together with their physical interactions and chemical reactions. Pathway model reconstruction is currently largely carried out manually by domain experts, who must carefully read the scientific literature, in order to retrieve, evaluate and interpret and distil relevant fine-grained statements. Moreover, due to the proliferation of scientific databases and ontologies, discovery of previously unknown knowledge demands that scientists take into account information from many different resources, covering different levels of contextual information (e.g., degree of confidence or certainty expressed towards a finding). Thus, given the high complexity mechanisms involved in pathway models, whose detailed description can only be derived from analysis of heterogeneous, fragmented and incomplete sources, reconstructing pathway models is a slow, difficult and laborious process. Accordingly, there is a need to develop methods that help experts to make sense of the continuously growing body of literature, in order to increase the speed and reliability of knowledge discovery.

In response to the above, text mining (TM) aims to automate the above process, by finding relations (such as interactions) that hold between concepts of different types (e.g., genes/proteins, chemical compounds, metabolites, subcellular components, anatomical entities, organisms, cell lines, strains, diseases). A large number of TM methods aim to extract simple binary relations from e.g., A binds B. This is mainly achieved by focusing on textual co-occurrences, using bag-of-words approaches, analysis of controlled vocabulary metadata, and other shallow techniques. However, these approaches have several disadvantages, including the identification of many false positive relations. Additionally, they fail to take into account contextual information about relations, e.g., the cellular context of a signaling event, such as cell type and localization.

In contrast, our work involves the development of more sophisticated TM techniques to extract events, which encapsulate typed n-ary relationships, i.e., interactions between any number of concepts. Events are able to capture detailed information about mechanisms of biological pertinence, e.g.,, reactions such as negative regulation, phosphorylation, carboxylation), by linking together interacting participants, which play specific roles (e.g., modifier, reactant, product, cause, location). As such, they are able to encode several types of contextual information, that are frequently missing when only binary relations are considered.

Consider an intuitive example from the literature to explain our goal: The results suggest that the narL gene product activates the nitrate reductase operon. (PMID: 3035558). This sentence provides interpretative information about the reaction between the narL gene product and the nitrate reductase operon, namely that the information stated in based on an analysis/interpretation of experimental results, and that there is a certain amount of speculation expressed towards the reaction (according to the use of the verb suggest, rather than a more definite verb, such as demonstrate). Next, consider a more complex example: The analysis showed that IEXC29S was unable to significantly transactivate the c-sis/PDGF-B promoter. Whilst a conventional TM analysis to find binary relationships would simply discover that some type of interaction occurs between IEXC29S and c-sis/PDFG-B, a more detailed contextual analysis would allow the construction of a representation that encodes the complex details of the interaction, e.g., that the information is stated based on an experimental analysis, and that the interaction has been shown to occur with a low level of intensity. .

In order to extract such complex events automatically, we have developed a pipeline-based event extraction system, EventMine [1], which employs a series of classifier modules to capture core event elements: detection of triggers (words or phrases that characterise the event; typically verbs or their nominalisations ,detection of edges (finding links between pairs of concepts), and complex event detection (combining multiple edges of complex n-ary relations).

EventMine utilises a rich set of features including those obtained from dependency parse trees supplied by the GENIA Dependency Parser [2], as well as from predicate-argument structures determined by Enju [3], which has been adapted for application to biomedical text. EventMine is capable of extracting interactions across different sentences, owing to its capability to incorporate results from a pre-executed coreference resolution method [4]. In this way, event participants

Proceedings of the XVIII International Conference «Data Analytics and Management in Data Intensive Domains» (DAMDID/RCDL'2016), Ershovo, Russia, October 11 - 14, 2016

which are semantically empty (e.g., expressions such as it, that) are resolved to their referents and thus become more informative. In addition, the system can be adapted to different tasks without the need for task-specific tuning [5]. Finally, EventMine also facilitates the extraction of interpretative context by detecting various event attributes, e.g., polarity, certainty, manner, knowledge type and source [6]. As with its other classifier modules, EventMine uses SVMs for this task, facilitated through its training on the GENIA Meta-knowledge corpus [7].

The automatic extraction of events from biomedical text has a broad range of applications, which include not only support for the creation and annotation of pathways [8], but also automatic population/enrichment of databases and semantic search systems. To develop systems that are customized for different tasks such as the above, a text mining infrastructure is needed, which is able to support the curation and maintenance of pathways, sharing and re-using of knowledge distributed over thousands of scientific publications and monitoring of recent publications is needed to maintain relevance. To foster adaptability of TM solutions, we are using our UIMA based Argo platform [9], which enables the development of highly customisable solutions in the form of reconfigurable modular text mining pipelines (workflows). Apart from supporting the straightforward integration of application-specific components, reconfigurable workflows allow for discovery of optimal solutions [10] owing to their interchangeable underlying components. For components to be interoperable (i.e., for one component to build on the text annotations created by another), the outputs of a predecessor component must be type-compatible with the inputs expected by a successor. In Argo, this is ensured by mechanisms that support mapping between similar semantic types and conversion of annotations [11].

Argo has supported the development of systems such as PathText2[1], an integrated search system that links biological pathways with supporting knowledge in the literature [8]. It reads formal pathway models (represented in the Systems Biology Markup Language (SBML) and converts them into queries that are submitted to three semantic search systems operating over MEDLINE, i.e., KLEIO, which improves and expands on standard literature querying with semantic categories and facetted search, FACTA+ (see below) and MEDIE (http://www.nactem.ac.uk/medie/), which extracts events. MEDIE has been found to achieve the highest hit ratio, which demonstrates the superiority of events for finding relevant interactions.

FACTA+[2] [12] discovers hidden, previously unknown associations between both concepts and complex events from the literature (such as Gene expression, Positive regulation, Binding, Regulation, etc.). Such associations can often only be found by linking information that may be dispersed across many documents, and thus which might be missed using manual search methods. This facilitates hypothesis generation, which is directly relevant to pathway construction. FACTA+ approaches the problem of automatic discovery of useful hypotheses by combining two (or more) known associations, i.e., if concept X is associated with concept Y and concept Y is associated with concept Z, then the potential indirect association between X and Z is considered as a useful hypothesis unless there is already a known association between X and Z. FACTA+ supports the discovery of indirect associations based not only on concepts but also on complex events such as Gene expression, Positive regulation, Binding, Regulation, etc.

Advanced TM methods such as those described here support pathway curation, validation and maintenance. Their employment yields improved coverage, faster acquisition and throughput, combined with easier identification and normalisation of duplicates, greater consistency, completeness and accuracy in description, and reduced curator burden. This helps to free experts from mundane and tedious tasks while aiding with more intellectually challenging ones.

References

[1] Miwa M, Saetre R, Kim JD, Tsujii J. Event extraction with complex event classification using rich features. J Bioinform Comput Biol. 2010;8(1):131-46. Epub 2010/02/26. doi: S0219720010004586 [pii]. PubMed PMID: 20183879.

[2] Sagae K, Tsujii Ji. Dependency parsing and domain adaptation with LR models and parser ensembles. Proceedings of CoNLL 2007 Shared Task; 2007. p. 1044-50.

[3] Miyao Y, Sagae K, Saetre R, Matsuzaki T, Tsujii J. Evaluating contributions of natural language parsers to protein-protein interaction extraction. Bioinformatics. 2009;25(3):394-400. PubMed PMID: 19073593.

[4] Miwa M, Thompson P, Ananiadou S. Boosting automatic event extraction from the literature using domain adaptation and coreference resolution. Bioinformatics. 2012;28(13):1759-65. doi: 10.1093/bioinformatics/bts237.

[5] Miwa M, Ananiadou S. Adaptable, high recall, event extraction system with minimal configuration. BMC Bioinformatics. 2015;16(10):1-11. doi: 10.1186/1471-2105-16-s10-s7.

[6] Miwa M, Thompson P, McNaught J, Kell DB, Ananiadou S. Extracting semantically enriched events from biomedical literature. BMC Bioinformatics. 2012;13(1):108.

[7] Thompson P, Nawaz R, McNaught J, Ananiadou S. Enriching a biomedical event corpus with meta-knowledge annotation. BMC Bioinformatics. 2011;12:393.

[1] http://www.nactem.ac.uk/pathtext2/demo/

[2] http://www.nactem.ac.uk/facta/

[8] Miwa M, Ohta T, Rak R, Rowley A, Kell DB, Pyysalo S, et al. A method for integrating and ranking the evidence for biochemical pathways by mining reactions from text. Bioinformatics. 2013;29(13):i44-i52.

[9] Rak R, Rowley A, Black W, Ananiadou S. Argo: an integrative, interactive, text mining-based workbench supporting curation. Database: The Journal of Biological Databases and Curation. 2012;2012.

[10] Batista-Navarro R, Rak R, Ananiadou S. Optimising chemical named entity recognition with pre-processing analytics, knowledge-rich features and heuristics. Journal of Cheminformatics. 2015;7(1):1-13. doi: 10.1186/1758-2946-7-s1-s6.

[11] Batista-Navarro, R., Carter, J. and Ananiadou, S. Argo: enabling the development of bespoke workflows and services for disease annotation. Database: the Journal of Biological Databases and Curation, 2016; doi:10.1093/database/baw66

[12] Tsuruoka, Y., Miwa, M., Hamamoto, K., Tsujii, J., and S. Ananiadou (2011) Discovering and visualizing indirect associations between biomedical concepts, in Bioinformatics, 27(13), i111-i119

Извлечение данных из текстов

Data Extraction from Texts

Извлечение низкочастотных терминов из специализированных текстов

© К. К. Боярский
Университет ИТМО
Санкт-Петербург
boyarin9@yandex.ru

© Н. А. Арчакова © Е. А. Каневский
Экономико-математический институт РАН
Санкт-Петербург
Assoul@yandex.ru kanev@emi.nw.ru

Аннотация

Исследована возможность повышения качества выделения терминов предметной области в узкоспециализированных научных текстах. Для этого вначале с помощью семантико-синтаксического анализа и построения дерева зависимостей выделялись тематически значимые фрагменты текста. Затем производились кластеризация фрагментов и поиск терминов с использованием семантического классификатора. Показано, что данный метод позволяет с высокой вероятностью обнаруживать термины даже с единичной встречаемостью.

1 Введение

Во многих дисциплинах сейчас разрабатываются стандартные онтологии предметных областей, которые предназначены для совместного использования экспертами и автоматическими системами обработки информации. Процесс создания онтологии характеризуется высокой трудоемкостью, поскольку необходимо адекватно и максимально полно описать каждый концепт (термин), входящий в нее, с указанием всех возможных связей с другими концептами. В нашем исследовании анализируется начальный этап построения онтологии предметной области – автоматическое формирование списка терминов. В качестве исходного материала были взяты статьи Большого экономического словаря [7]. Представлены результаты кластеризации корпуса определений экономических понятий с последующим выделением терминов для каждого кластера. Объем корпуса был ограничен размерами, позволяющими вести ручную экспертную обработку для контроля качества работы автомата.

Труды XVIII Международной конференции DAMDID/RCDL'2016 «Аналитика и управление данными в областях с интенсивным использованием данных», Ершово, 11-14 октября 2016

2 Смежные исследования

Отбор терминов для помещения в онтологию обычно осуществляется с помощью лингвистического и статистического анализа [3]. В работе [2] использован гибридный метод: сначала отбираются все существительные, из которых потом по результатам статистического анализа с использованием информации о семантике слов формируется список возможных терминов.

Для повышения точности отбора в [1] применялся кластерный анализ англоязычного текста, использующий адаптивный алгоритм Леска «с помощью нахождения пересечений смыслов слов в WordNet». Результаты применения метода сравнивались с результатами, полученными при использовании частотного словаря, частотного семантического словаря и позиционного метода. Стоит отметить, что в [1] изначально показатели полноты и точности в определении ключевых слов по частотному словарю были довольно высокими. Схожая методика используется и в нашей работе. Однако целью работы [1] было извлечения ключевых слов, характеризующих каждый текст из корпуса, а цель нашего исследования – извлечение терминов, характеризующих конкретную область знаний.

В [11] рассматривалась задача онтологического анализа терминологических словарей методом семантического анализа дефиниций и построения OWL-описаний выделенных объектов. Однако выявление концептов для описания в основном базировалось на ручной обработке.

3 Специфика текста и предварительная обработка

Для анализа мы выбрали фрагмент Большого экономического словаря [7], состоящий из 1020 словарных статей, относящихся к банковской деятельности.

Для контроля качества работы системы предварительно вручную было выделено 462 однословных термина. Мы сосредоточились именно на однословных терминах, потому, что их автоматическое выделение представляет наибольшие трудности. Двухсловные термины в данных текстах имеют четко выраженную структуру: прилагательное + существительное, расположенные контактно, и могут быть найдены стандартными частотными методами.

Выбранный тип текста (словарь) имеет свои особенности. С одной стороны, все словарные статьи построены по одному шаблону: заголовок статьи и дефиниция, состоящая, как правило, из гиперонима и дополнительной информации. При этом термины предметной области могут встречаться в обеих частях словарной статьи. Как оказалось, в заголовочные части входит только 59% терминов. Такая структурированность облегчает обработку текста.

С другой стороны, как отмечалось в [6], для текстов научной направленности типична ситуация, при которой одни термины предметной области встречаются очень часто (БАНК – 826 вхождений, 3% от всех неслужебных слов корпуса), а другие – имеют только единичные вхождения (РЕПРАТТА, ХЕДЖЕР и др.). При стандартных методах отбора лексем (типа TF-IDF, LDA) для дальнейшей обработки, отбрасываются как те, так и другие. В то же время среди слов со средней частотностью большую часть составляют слова общей лексики. Например, термин ПОКУПАТЕЛЬ с абсолютной частотой вхождений 45 в анализируемом корпусе имеет по частоте встречаемости существительных 56-й ранг из 98. В первом столбце табл. 1, построенному по частотному словарю, приведены 19 существительных из окрестности слова ПОКУПАТЕЛЬ в интервале с 50 ранга (частота 51) по 62 ранг (частота 39). Оказывается, что к экономическим терминам (выделены жирным шрифтом) относятся только 42% лемм.

Таблица 1 Сравнение разных методов создания частотных списков

Методы	**Частотный словарь**	**TF-IDF**	**Предлагаемый метод**
Леммы	условия	институт	**заемщик**
	покупка	соглашение	**кредитор**
	депозит	условия	**аккредитив**
	прибыль	**актив**	**цена**
	средства	**владелец**	средства
	требование	**прибыль**	**оплата**
	использование	уровень	**вкладчик**
	производство	обращение	**сделка**
	вложение	орган	**затрата**
	ПОКУПАТЕЛЬ	**ПОКУПАТЕЛЬ**	**ПОКУПАТЕЛЬ**
	Расход	**договор**	**компания**
	соглашение	**залог**	**облигация**
	договор	требование	**залог**
	вкладчик	часть	**валюта**
	часть	**долг**	**чек**
	владелец	осуществление	**фонд**
	обращение	**погашение**	организация
	риск	**чек**	**документ**
	время	время	**рынок**
Процент терминов	42%	47%	89%

Аналогичные результаты получаются при анализе распределения терминов по TF-IDF (табл. 1, второй столбец). Здесь термин ПОКУПАТЕЛЬ имеет 38-й ранг из 80. В таком же диапазоне из 19 существительных (относящихся к 36–43 рангам) терминами являются 47% лемм.

Дополнительные проблемы создает использование модели bag-of-words, явно противоречащей структурности словарной статьи. При проведении данного исследования вместо bag-of-words использовался фрагмент дерева подчинения, построенного с помощью семантико-синтаксического парсера SemSin [8]. Этот фрагмент («краткое определение») включал в себя заголовочный термин и его гиперонимы с препозитивными определениями и зависимыми существительными в родительном падеже (рис. 1).

В качестве примера рассмотрим следующую словарную статью (слова, включенные в краткое определение, выделены жирным шрифтом):

*ОБМЕННЫЙ КУРС – **курс**, по которому одна валюта обменивается на другую, **цена денежной единицы** страны, выраженная в иностранной валюте <…>.*

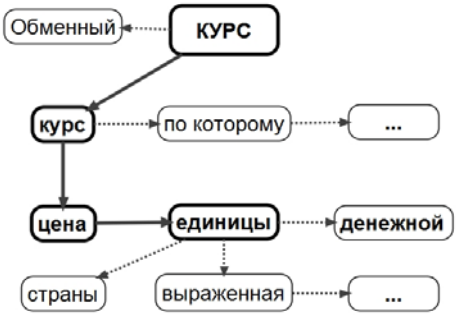


Рисунок 1 Построение краткого определения по фрагменту дерева подчинения

В правую часть кратких определений входит 16% выделенных вручную терминов, остальные – в полные определения. Такой подход позволил исключить из анализа большое количество слов

общей лексики и выявить «нелокальные» конструкции, состоящие из неконтактно расположенных слов

4. Кластеризация словарных статей

Выделение терминов происходило в три этапа:

1. Кластеризация текста.
2. Выделение наиболее частотных классов для каждого кластера.
3. Формирование списка терминов из слов, принадлежащих выделенным классам.

Для сравнения дефиниций была построена векторная модель текста. Каждому *i*-му абзацу, соответствующему одной словарной статье, был приписан нормализованный вектор $\{w_{n,i}\}$, где $w_{n,i}$ – частота токена *n* в *i*-ом абзаце. В качестве токена выступала либо лемма-существительное (сравнение «по леммам»), либо ее семантический класс (сравнение «по классам»), выявленный в ходе синтаксического анализа, выполняемого парсером SemSin. В последнем случае предполагалось, что слова, принадлежащие одному классу, эквивалентны: например, леммы *банкнота* и *валюта* относятся к одному классу «Купюры». Классы определялись в соответствии с семантическим классификатором, аналогичным описанному в [13]. В нем 192 тыс. лексем распределены по 1688 классам.

Существует много разных способов кластеризации данных. Выбор наиболее подходящего способа обусловлен особенностями анализируемых данных и целей исследования. Был применен стандартный иерархический агломеративный алгоритм кластеризации Уорда. Метод Уорда направлен на минимизацию суммы разностей квадратов расстояний внутри каждого кластера [5, 9].

Для кластеризации мы использовали пакет scikit-learn (Python) [4]. Данная реализация алгоритма накладывает ограничения на выбор способа измерения расстояния между векторами, поэтому была применена метрика Евклида.

Сначала все объекты являются отдельными кластерами. На каждой итерации алгоритма к текущему кластеру *s* добавляется один объект *t* так, что межкластерное расстояние между новым кластером $u = s \cap t$ и любым другим кластером меньше внутрикластерного расстояния.

Во многих отношениях метод Уорда является наиболее точным среди других иерархических методов [5]. В отличие от неиерархических методов, метод Уорда является устойчивым (не зависит от выбора начального приближения) и выделяет кластеры произвольной формы. Кроме того, как указано в [9], расстояние по Уорду обладает свойством растяжения. Это означает, что по мере роста кластера, расстояние от него до остальных кластеров увеличивается, что приводит к более чистому результату даже для «низкоконтрастных» текстов, в которых переход к новому разделу не означает смены лексикона. К недостаткам иерархического метода кластеризации относится то, что один из образуемых кластеров, как правило, значительно больше всех остальных.

Оптимальным для дальнейшей обработки оказалось разбиение текста на 35 кластеров, включающих от 4 до 282 словарных статей (среднее значение равно 29, медиана — 18).

Варианты классификации «по классам» и «по леммам» сравнивались с использованием данных о внутрикластерных и межкластерных расстояниях (Табл. 2). Чем больше разница между этими параметрами, тем точнее проведена кластеризация. Для используемой метрики расстояние лежит в интервале 0…1,41.

Таблица 2 Средние значения внутрикластерных и межкластерных расстояний

Вариант отбора лексики	**среднее внутрикластерное**	**среднее межкластерное**
«по классам»	0.41	1.04
«по леммам»	0.45	1.0

Отметим, что результаты кластеризации подтверждают вывод, сделанный в [6]: основной классифицирующей силой в русскоязычных текстах обладают существительные. Если при кластеризации по леммам-существительным отношение межкластерного и внутрикластерного расстояния равно 2,2 (табл. 2), то в контрольной кластеризации с учетом также прилагательных и глаголов оно составило только 1,08.

В целом, вариант отбора лексики «по классам» показывает более точные результаты. Например, термин ЦЕССИОНАРИЙ имеет три значения: лицо, становящееся кредитором (1); правопреемник (2); страховая компания (3). Во всех вариантах значение (1) верно определяется как относящееся к кластеру «Люди». При сравнении «по леммам», значения (2) и (3) объединяются в кластер, состоящий из 436 словарных статей. При сравнении «по классам» эта лемма в значении (2) попадает в кластер «Люди», а в значении (3) в кластер «Финансовые организации».

Для более точного исследования разницы результатов отбора лексики «по леммам» и «по классам», было сформировано два списка: список терминов-кандидатов «по классам» и список терминов-кандидатов «по словам» соответственно.

5 Автоматическое выделение терминов

Кластеризация, описанная в разделе 4, проводилась на основе словаря кратких определений, что позволило существенно понизить зашумленность данных. На следующих этапах алгоритма выделения

терминов мы использовали изначальные полные словарные статьи, поскольку, как было отмечено ранее, термины могут встречаться в любой части словарной статьи.

Сначала было выделено по три наиболее частых класса для каждого кластера. Стоит отметить, что полученный список классов, отличается от частотного списка классов, созданного до кластеризации из данных обо всем тексте в целом. Например, при отборе лексики «по классам» после кластеризации найдено 36 самых распространенных классов (Финансы, Деньги, Платежи, Учреждения…). Кроме того, обнаружено 7 существенных классов, которых не было в изначальном частотном списке классов (Документы, Торговля и сервис…). К ним относятся, например, термины РЫНОК, АУКЦИОН.

Затем, для каждого кластера отбирались существительные, принадлежащие наиболее частотным классам данного кластера. Эти существительные и вошли в итоговый список терминов-кандидатов.

6 Обсуждение результатов

6.1 Анализ списка терминов-кандидатов «по классам»

Из 311 отобранных слов «по классам» к терминам, выделенным экспертами относилось 249. Таким образом, точность отбора составила 0,8. В третьем столбце табл. 1 показана окрестность термина ПОКУПАТЕЛЬ (19 ранг) в частотном словаре итогового отбора. Если, как и выше, рассмотреть группу из 19 слов (с 13 по 24 ранги), то оказывается, что доля терминов в выборке увеличилась до 89% (по исходному частотному словарю – 42%). При этом из 17 терминов 13 «уникальны», т. е. встречаются в корпусе только один раз, например, ЗАЕМЩИК, ВАЛЮТА, ФОНД, РЫНОК. В то же время, в остальных столбцах присутствуют только по три уникальных термина.

Лексический анализ показал, что итоговый список терминов-кандидатов включает 95 из 147 терминов с единичной встречаемостью: ПРЕДОПЛАТА, ЛУИДОР, ФИДУЦИАРИЙ, КОМПАНИЯ-ХОЛДИНГ…. Из них 40 терминов встречались только в правой части словарной статьи, включая 25 терминов, не вошедших в краткие определения.

Для оценки качества выделения терминов был выбран стандартный способ оценки эффективности выделения информации, основанный на показателях точности и полноты. В нашем случае точность – доля правильных терминов среди слов-кандидатов, найденных автоматически, а полнота показывает, какую часть из терминов, выделенных экспертами, удалось обнаружить автоматически.

На рис. 2 и 3 представлены оценки точности и полноты соответственно. Предварительно список терминов-кандидатов был разбит на 10 примерно равных интервалов так, что термины-кандидаты с одинаковой частотой были в одном интервале. По оси абсцисс показана округленная средняя частотность терминов-кандидатов в данном интервале.

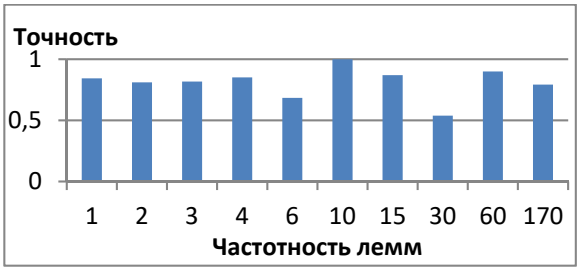


Рисунок 2 Зависимость точности выделения терминов от частотности их употребления в тексте

Как видно из рис. 2, точность не зависит от частоты встречаемости термина-кандидата. Между тем, чем чаще встречается термин, тем выше вероятность, что он будет обнаружен автоматически (рис. 3). Стоит заметить, что представленный метод извлекает термины, встречающиеся один или два раза, с вероятностью 40%. Данные из таблицы 3 показывают, что стандартные методы выделяют такие низкочастотные термины значительно хуже.

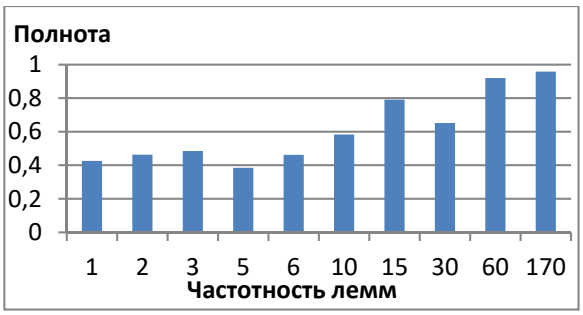


Рисунок 3 Зависимость полноты выделения терминов от частотности их употребления в тексте

В табл. 3 представлены средние оценки точности и полноты, подсчитанные с помощью 4-х разных методов:

1. по частотному словарю;
2. по списку слов, составленному по весам TF-IDF;
3. по частотному словарю слов, являющихся заголовками словарных статей;
4. по частотному списку терминов-кандидатов, созданному по описанному методу «по леммам» и «по классам».

Количество слов во всех списках одинаковое (311 по количеству терминов-кандидатов, выделенных автоматически), слова располагаются в порядке убывания частоты встречаемости. Например, частотность в первом списке изменяется от 826 до 10. Каждый список сравнивался со списком терминов, выделенных экспертами. По таблице 3 видно, что результаты предлагаемого метода значительно выше

остальных представленных способов выделения терминов.

Таблица 3 Точность и полнота, полученные по разным методам выделения терминов

	Точность	Полнота	F-мера
Частотный словарь	0.37	0.25	0.3
TF-IDF	0.27	0.19	0.22
Частотный словарь по заголовкам	0.63	0.42	0.5
Предлагаемый метод «по леммам»	0.64	0.48	0.55
Предлагаемый метод «по классам»	0.8	0.54	0.64

Был проведен анализ терминов-кандидатов, не отмеченных экспертами как термины. Можно выделить несколько причин избыточности списка терминов-кандидатов:

1. эти слова являются частью многословных терминов: ресурсы (финансовые ресурсы), карточка (пластиковая карточка), бумага (ценная бумага);
2. это гипероним, относящийся к общей лексики (организация, лицо);
3. эти слова описывают семантическую иерархию (сеть, множество, объединения, вид).

6.2. Сравнение списков терминов-кандидатов «по классам» и «по леммам»

Список терминов-кандидатов «по леммам» включает 349 лемм. Как видно из таблицы 3, результаты оценки качества выделения терминов немного выше результатов выделения по заголовкам, но значительно ниже результатов «по классам».

Списки терминов-кандидатов «по классам» и «по леммам» включают 192 общих терминов (42% от общего количества терминов): ВАЛЮТА, ДОЛЛАР, ЧЕКОДЕРЖАТЕЛЬ… и 54 слова общей лексики: БУМАГА, ГАРАНТИЯ, ОРГАНИЗАЦИЯ, ОСНОВАТЕЛЬ.… При этом 57 терминов встречаются только в списке терминов-кандидатов «по классам»: ФИНАНСЫ, БРОКЕР, ВАРРАНТ, ДЕБЕТ… В списке терминов-кандидатов «по леммам» таких слов всего 30.

Таким образом, сравнение двух способов отбора первичных данных показало, что предпочтительнее использовать способ отбора «по классам».

7 Заключение

В работе представлены результаты многоэтапного выделения однословных терминов из словаря предметной области с использованием классификатора. Показано, что традиционные методы определения терминов по частотному словарю или по весам TF-IDF не дают верной картины распределения терминов в узкоспециализированном тексте. Кроме того, несмотря на очевидность решения задачи — выборки терминов из заголовков статей словаря — оказалось, что уникальные термины могут встречаться также в любых частях дефиниций. Качество выделения терминов-кандидатов оценивалось с помощью списка терминов, найденных экспертами.

Метод, примененный в этой работе, увеличивает вероятность выделения терминов почти в два раза. Он зависит не от начального частотного распределения терминов в тексте, а от качества кластеризации. Благодаря этому, учитываются как термины с единичным вхождением, так и высокочастотные термины, равномерно распределенные по всему тексту. Средняя по частотности точность выделения терминов составила 0,8, полнота – 0,54, F-мера – 0,65.

Для проверки общности полученных результатов была проведена обработка текстов совершенно иной структуры и из другой предметной области, а именно, глав двух монографий, посвященных парусному вооружению судов [10, 12]. Несмотря на то, что лексика этих книг сильно различается (между ними полтора века), общие закономерности сохраняются. Вручную было выделено 244 термина, имеющих отношение к кораблям. Как и в экономической сфере имеется термин с очень высокой частотностью (ПАРУС – 238 вхождений, 4,4% от всех неслужебных слов). Наряду с этим 112 терминов (46%) встречаются только один раз.

При обработке по описанному выше алгоритму было найдено 158 лемм-кандидатов, из которых 152 термина. Таким образом, точность выделения терминов для судовой тематики составила 0,96, а полнота – 0,62. Этот результат показывает независимость предлагаемого метода выделения терминов от выбора конкретной предметной области.

Литература

[1] Haggag M. H., Abutabl A., Basil A. Keyword extraction using Clustering and Semantic Analysis. International Journal of Science and Research, Vol.3 #11 (November 2014). Pp 1128-1132.

[2] Lacasta J., Nogueras-Iso J., Zarazaga-Soria J. Terminological Ontologies: Design, Management and Practical Applications. Semantic Web and Beyond: Computing for Human Experience. Springer-Verlag, 2010.

[3] Pazienza M. T., Pennacchiotti M., Zanzotto F.M. Terminology extraction: an analysis of linguistic and statistical approaches. Knowledge Mining, Springer Verlag, 2005. Pp 255-281.

[4] SciPy [Электронный ресурс] URL: http://docs.scipy.org/doc/scipy/reference/generated/scipy.cluster.hierarchy.linkage.html#scipy.cluster.hierarhie.linkage (дата обращения: 5.05.2016).

[5] Srinivasan R., Pepe A., Rodrigez V.F. A comparision between ethnographic and clustering-based semi-automatic technics for cultural ontologies. Journal of the American Society for Information Science and Technology, 2008, №5.
[6] Артемова Г.В., Боярский К.К., Гусарова Н.Ф., Добренко Н.В., Каневский Е.А. Категоризация текстов для структурирования массива исторических документов // Труды XVI Всероссийской научной конференции RCDL-2014 «Электронные библиотеки: перспективные методы и технологии, электронные коллекции». Дубна, 2014. С. 159–164.
[7] Борисов А.Б. Большой экономический словарь. – М.: Книжный мир, 2003. 895 с.
[8] Боярский К.К., Каневский Е.А. Семантико-синтаксический парсер SemSin // Научно-технический вестник информационных технологий, механики и оптики. 2015. т. 15. № 5. С. 869–876.
[9] Воронцов К.В. Лекции по алгоритмам кластеризации и многомерного шкалирования. 2007. URL: http://www.ccas.ru/voron/download/Clustering.pdf (дата обращения: 5.05.2016.
[10] Курти О. Постройка моделей судов. Энциклопедия судомоделизма. Сокращенный пер. с итал. А.А. Чебана. Л.: Судостроение, 1977. 544 с.
[11] Лезин Г.В., Клименко Е.Н., Силина Е.Ф. Онтологическая интерпретация дефиниций терминологического словаря // Прикладная лингвистика в науке и образовании. Сборник трудов VII международной конференции. – СПб.: «Книжный дом», 2014. С. 50–54.
[12] Ромм Ш. Морское искусство или Главные начала и правила, научающие искусству строения, вооружения, правления и вождения кораблей. Часть 1. Пер. с франц. А.А. Шишков. Типография Морского шляхетского кадетского корпуса. Часть 1, 1793. 542 с. Часть 2, 1795. 355 с.
[13] Тузов В.А. Компьютерная семантика русского языка. СПб: Изд-во С.-Петерб. ун-та, 2004

Extraction of low-frequent terms from domain-specific texts

K. Boyarsky, N. Archakova, E. Kanevsky

We examined the way to improve the quality of low-frequent term recognition in scientific texts. Firstly, domain-relevant fragments were extracted from the text with the help of dependency tree. Then the fragments were clustered and candidate terms were defined using the semantic classifier. The studies suggest that this approach allows extracting unique terms as well.

О применении дополнительных индексов часто встречающихся слов для полнотекстового поиска

© А. Б. Веретенников

Уральский федеральный университет,
Екатеринбург
alexander@veretennikov.ru, AlexanderBorisovich@urfu.ru

Аннотация

Разные слова в текстах встречаются с разной частотой. Рассматриваются дополнительные индексы, предназначенные для ускорения выполнения поискового запроса, который включает в себя часто встречающиеся слова. Вводится несколько групп слов, для которых разработаны разные методы. Рассмотрен вопрос оптимизации записи индекса.

1 Введение

Рассматривается задача полнотекстового поиска, то есть поиска фраз или наборов слов в текстах на естественном языке, например, русском или английском. Изучаем задачу поиска с учетом расстояния, в документах искомые слова должны располагаться как можно ближе друг к другу. Эта задача требует сохранения в индексе информации о каждом вхождении в документе каждого слова.

Слова в текстах встречаются с разной частотой. Считается, что распределение частот слов соответствует закону Ципфа [8]. Слова, которые часто используются, могут встречаться в тысячи и более раз чаще, чем редко используемые слова.

Скорость выполнения поискового запроса определяется наиболее часто встречающимися словами, входящими в него.

Разделим слова на группы, на основании их частоты встречаемости. Для разных групп слов определим разные методы их обработки, например, создание дополнительных индексов.

В [10] мы определили три группы слов:

1. «Стоп слова»: и, в, или. Встречаются очень часто и могут вообще не включаться в индекс, поэтому их и можно называть стоп словами. Например, предлоги. Далее будем назвать данные слова стоп словами, даже если в каком-то виде включаем информацию о них в индекс.
2. Часто используемые слова. Встречаются часто, но несут в себе существенный смысл и должны включаться в индекс.
3. Остальные, будем называть их «обычные слова».

Далее учитываем все виды слов при поиске, то есть, для любого слова информация в каком-то виде включается в индекс. В поисковом запросе учитываются все слова.

2 Морфологический анализатор

Морфологический анализатор для каждой словоформы возвращает список номеров базовых форм, где номер это число в промежутке [0, *WordsCount* – 1], где *WordsCount* – число базовых форм в словаре, около 260 тыс. для используемого словаря, описывающего русские и английские слова.

Базовые формы слова также называются леммами, а сам процесс получения набора лемм по словоформе – лемматизацией

С учетом морфологического анализа разделение слов на три группы при использовании анализатора применяется не к исходным словоформам, а к леммам. Соответственно, есть три типа лемм в смысле частоты встречаемости: стоп леммы, часто используемые леммы и остальные.

Если мы упорядочим леммы по частоте встречаемости, по убыванию, то вначале идет группа стоп лемм, затем группа часто используемых лемм, затем остальные леммы. Размеры групп определяются параметрами, в зависимости от задачи и способа создания дополнительных индексов.

3 Инвертированные индексы

Для решения задач полнотекстового поиска применяются инвертированные файлы [4] и их аналоги. Индекс состоит из двух частей.

Инвертированный индекс представляет собой набор записей вида (*ID*, *P*), *ID* – идентификатор документа, *P* – позиция, например порядковый номер, слова в документе. Запись (*ID*, *P*) будем называть записью о вхождении слова в документе или словопозицией. Все записи, соответствующие одной лемме, хранятся последовательно для их быстрого чтения при поиске. Идентификатор

Труды XVIII Международной конференции DAMDID/RCDL'2016 «Аналитика и управление данными в областях с интенсивным использованием данных», Ершово, 11-14 октября 2016

документа *ID* будем считать целым числом, например, номером документа.

Таблица дескрипторов. Содержит для каждой леммы структуру, дескриптор, которая содержит информацию о том, где в инвертированном индексе находятся данные для этой леммы.

Индекс это ассоциативный массив, в качестве ключа может выступать, например, лемма, в качестве значения – список словопозиций.

4 Поиск фраз в текстах

4.1 Существующие методы

В [1, 7] приводятся алгоритмы создания дополнительных индексов для более быстрого поиска фраз. Авторы предлагают две оптимизации:

1. Вводится частичный индекс фраз, в котором ключами являются часто используемые фразы (набор которых определяется часто используемыми запросами пользователей).
2. Вводится индекс пар слов (nextword – индекс). Для пары слов хранится информация об их вхождениях в текстах. При этом фиксируется порядок слов в паре и слова находятся непосредственно рядом друг с другом.

В [7] дается обоснование необходимости учета стоп слов при поиске. Простое устранение стоп слов из индекса и поиска может привести к непредсказуемым результатам, если стоп слово имеет для конкретного случая особый смысл (например, см. пример далее).

При этом на поиск накладываются условия:

1. Учет порядка слов фразы.
2. Отсутствие «лишних» слов в составе фразы в искомых текстах.

То есть ищутся только те тексты, в которых фраза встречается точно так, как она введена пользователем. Например, если в тексте есть фраза «Time and a world by Yes», то при поиске

1. Time and a world Yes,
2. Yes time and a world,

данный текст не будет найден методами [1, 7], по причине наличия «by» в тексте (для обоих запросов) и несовпадении порядка слов (для второго запроса). Примечание. Yes – название группы, «Time and a world» – название произведения.

В [2, 3] дано развитие идей nextword – индекса. Идет переход от пар слов к последовательностям из нескольких слов. Однако указанные недостатки [1, 7] остаются. Таким образом, методы из [1, 2, 3, 7] ограничены в применении.

4.2 Отличие предлагаемых методов от существующих методов поиска фраз

Существенное отличие настоящей работы в том, что мы ищем без учета порядка слов, и поиск осуществляется с учетом того, что в текстах посередине фразы может еще что-то быть. Поэтому мы и говорим о поиске «**наборов слов**» и фраз.

Предлагаемые нами методы предлагают принципиально более качественный уровень решения проблемы и их потенциальная область применения существенно шире. На поисковый запрос не накладывается никаких дополнительных условий. Поиск с использованием дополнительных индексов по своим параметрам идентичен поиску с использованием обычных индексов, но выполняется существенно быстрее. За исключением двух ограничений. Во-первых, слова фразы в искомых текстах должны быть все-таки вблизи, то есть число «лишних» слов между ними должно быть ограничено (допустимое количество настраивается параметром). Во вторых, для запросов, которые состоят *только* из стоп лемм, не допускаем в тексте лишних слов во фразе, но ищем без учета порядка.

5 Обработка стоп слов

5.1 Обработка пар стоп лемм и стоп лемм, располагающихся радом с другим словом

В [10] рассмотрен простой способ обработки стоп лемм. Во первых, создается дополнительный индекс, ключ в котором – пара стоп лемм, для которого в индексе хранятся позиции в документах, где эти стоп леммы находятся непосредственно рядом.

Во вторых, если мы имеем словопозицию некоторого слова, то в ее состав включается информация о находящихся непосредственно рядом с этим словом стоп леммах. Алгоритм кодирования описан в [10].

Данный подход позволяет ускорять поисковые запросы, в которых рядом с обычными словами находятся стоп леммы, а также поисковые запросы, включающие в себя нескольких стоп лемм.

5.2 Обработка фраз, состоящих из стоп слов

Если мы создаем индекс для пар стоп лемм, а в поисковом запросе их больше, то использовать пары неудобно. К примеру, если запрос включает в себя n стоп лемм, то может потребоваться обработать порядка $n*n$ пар стоп лемм.

В [11] рассмотрен индекс, ключ в котором определяется набором стоп лемм. Если в тексте встречается последовательность идущих друг за другом стоп лемм некоторой длины (заданы ограничения снизу и сверху, на длину последовательности), то формируется ключ, на основании этого набора стоп лемм. Словопозиция, соответствующая первой из них, включается в индекс. Набор лемм при создании ключа сортируется, поэтому поиск осуществляется без учета порядка.

Это позволяет ускорить поиск фраз, состоящих только из стоп лемм, длиной более 2-х слов.

Кроме того, для словопозиции произвольного слова рассмотрен вариант хранения в основном индексе информации о леммах стоп слов, которые

находятся на расстоянии от данного слова не более чем *MaxDistance* (параметр, например, 5).

Такие подходы имеют смысл исходя из:

1. Вследствие того, что стоп слова очень часто встречаются, практически для каждой их комбинации есть вхождение в текстах.
2. Существенная часть запросов, в которые входят стоп слова, это составные термины или часто используемые фразы, в которых состав слов зафиксирован.
3. Один из видов запросов, включающих стоп слова, это фрагмент текста, который получен путем копирования фразы из уже существующего текста, для поиска других документов, включающих подобную фразу.

6 Дополнительные индексы часто используемых слов

В [10, 11] расширенный индекс (*w*,*v*) это список вхождений слова *w*, когда в тексте не более чем на расстоянии *ProcessingDistance* от *w* присутствовало слово *v*. С учетом морфологического анализа под *w* и *v* мы понимаем леммы слов. Лемма *w* является часто используемой, лемма *v* – произвольная.

Параметр *ProcessingDistance* может быть задан различным в зависимости от частоты встречаемости леммы *w* в текстах. В экспериментах использовалось значение *ProcessingDistance* от 5 до 7.

Мы считаем, если расстояние между словами меньше или равно *ProcessingDistance*, то слова связаны по смыслу друг с другом, иначе нет.

В [11] описаны примеры применения дополнительных индексов для стоп базовых форм и часто используемых базовых форм для выполнения запросов, включающих разные комбинации разных видов слов.

В [11] также представлен алгоритм создания индекса стоп слов и результаты экспериментов, показывающие, что применение дополнительных индексов позволяет значительно (более чем в 10 раз) ускорить выполнение ряда запросов.

В [12] представлен алгоритм создания расширенных индексов и результаты экспериментов построения таких индексов.

7 Примеры выполнения поискового запроса

Рассмотрим запрос: "every stick has two ends".

В данном запросе stick, end, every – часто используемые слова, остальные слова – стоп слова. Каждое слово имеет одну базовую форму.

Для выполнения запроса с использованием дополнительных индексов требуется прочитать три списка словопозиций:

1. stick – обычный индекс,
2. (stick, end) – расширенный индекс,
3. (stick, every) – расширенный индекс,

При этом информация о стоп леммах has, two будет извлекаться из потока словопозиций stick. Для каждой словопозиции *T* в нем хранится информация о стоп леммах, что находятся в тексте близко от *T* (расстояние не более *MaxDistance,* в словах*)*.

В списке словопозиций (stick, end) элементов значительно меньше, чем в списке словопозиций для леммы (end), за счет чего получаем ускорение при выполнении поискового запроса.

Заметим, что запросы рассматриваются не как фразы, а как наборы слов. К примеру, если в тексте есть «every stick has two ends», то запрос «every stick two ends» также найдет это вхождение.

Рассмотрим запрос: "A thing of beauty is a joy for ever". В данном запросе thing, joy, ever – часто используемые слова, beauty – обычное слово, остальные слова – стоп слова. Каждое слово имеет одну базовую форму.

Для выполнения запроса с использованием дополнительных индексов требуется прочитать четыре списка словопозиций

1. beauty – обычный индекс,
2. (beauty, thing) – расширенный индекс,
3. (beauty, ever) – расширенный индекс,
4. (beauty, joy) – расширенный индекс.

8 Об алгоритме поиска

Мы можем обрабатывать запросы, включающие все виды слов. Запрос это несколько слов. В данном разделе мы рассмотрим одну из подзадач поиска, а именно, обработку поискового запроса, в котором:

1. У слов запроса нет стоп лемм.
2. Хотя бы для одного слова запроса все леммы часто используемые.

Будет предложен не рассмотренный ранее алгоритм поиска. Смысл алгоритма: выберем некое слово *a* в запросе, такое, что все леммы *a* часто используемые, назовем его основным словом запроса. Для каждого другого слова *b* рассмотрим расширенные индексы (*a*, *b*). Расширенный индекс содержит словопозиции слова *a*, когда *b* было близко с *a*. Рассмотрим последовательно каждую позицию слова *a* в текстах, и проверим, находились ли близко с *a* все из требуемых слов запроса.

То есть, если есть словопозиция (*ID1*, *P1*) слова *a*, то для каждого другого слова запроса *b* расширенный индекс (*a*, *b*) должен содержать словопозицию (*ID1*, *P1*). Рассмотрим данный подход с учетом того, что словоформа может иметь несколько лемм.

8.1 Определения

Секцией будем называть объект, который содержит в себе лемму и список словопозиций. Поля секции:

Lemma. Лемма.

Postings. Список словопозиций.

Value. Определяет текущий элемент списка словопозиций. Вначале это первый элемент списка,

затем мы можем последовательно брать следующий элемент списка.

Введем понятие «клетка» поискового запроса. Клеткой будем называть список секций. Поисковый запрос в структурированном виде это список клеток.

Секция является итератором словопозиций, то есть имеет операцию *Next*, возвращающую следующую запись. Реализация итератора включает в себя чтение списка словопозиций из индекса. Вначале итератор соответствует первой записи списка словопозиций. Операция *Next* позволяет перейти к следующей записи.

Клетка в свою очередь является итератором словопозиций. Итераторы секций образуют упорядоченный список. Итератор с минимальным значением текущей словопозиции находится в начале списка и его значение есть значение итератора клетки *Value*. При выполнении *Next* для клетки используем *Next* текущей секции, которая затем перемещается в списке, если ее значение становится больше, чем значение другой секции.

Запрос «двести сорок миль» соответствует трем клеткам: [двести] [сорока, сорок] [миля]. Слово «сорок» имеет две леммы, остальные слова по одной лемме.

Выберем индекс *M*, который соответствует клетке запроса, с минимальной суммарной частотой встречаемости лемм, среди таких клеток, у которых все леммы часто используемые. Для каждой секции *w* клетки *M*, и для каждой секции *v* остальных клеток существует расширенный индекс (*w*, *v*). Этот расширенный индекс определит список словопозиций *Postings* для секции *v*.

Будем считать, что для *v* есть список словопозиций *Postings,* состоящий из записей вида (*P*, *Delta*, *DeltaFlag*), где:

- *P* – позиция (номер) слова *w* в документе,
- *Delta* – расстояние от *w* до *v*,
- *DeltaFlag* – определяет, до или после *w* находится *v* в тексте.

Список *Postings* упорядочен по полю *P*, по возрастанию. В данном списке нет *ID* документа, потому что алгоритм применяется в рамках обработки одного документа. После обработки одного документа списки словопозиций очищаются, и заполняются данными следующего документа, затем поиск повторяется для него.

8.2 Вспомогательные переменные

Введем набор массивов, количество элементов в которых равно *L* = («Число клеток запроса» – 1). Одна ячейка массива соответствует одной клетке запроса, не совпадающей с основной клеткой *M*.

Flags. Битовый массив.

Positions, PositionFlags. Массивы расстояний. Пусть ячейка массива соответствует клетке *T*. В тексте есть лемма *w* клетки *M* и лемма *v* клетки *T*. В ячейке *Positions* хранится расстояние от *v* до *w*, а в *PositionFlags* – до или после *w* находится *v* в тексте.

Lemmas. Каждый элемент массива содержит список идентификаторов лемм клетки запроса.

Дополнительные переменные:

Current. Текущая позиция.

Marked. Счетчик тех клеток запроса *V*, леммы которых были найдены, т. е. выполнялось *V.Value.P* = *Current*.

8.3 Алгоритм

Формируем временную клетку *S*, в которую помещаем все секции клеток, не совпадающих с *M*. Выполняем в цикле:

1. Если *S.Value.P* не равно *Current* то:
   a. Если *Marked* = *L,* то мы нашли искомое. Сохраняем позицию *Current* в результатах поиска.
   b. Очищаем *Flags* (все ячейки теперь равны 0).
   c. Присваиваем *Marked* = 0.
   d. *Current* = *S.Value.P*.
2. Помечаем S.Lemma. Выполняем одно из двух.
   a. Если есть индекс i, такой что Lemmas[i] содержит S.Lemma и Flags[i] = 0, то: меняем Flags[1] на 1, сохраняем S.Value.Delta в Positions[i], PositionFlags[i] = S.Value.DeltaFlag, Marked = Marked + 1.
   b. Если есть индекс i, такой что Lemmas[i] содержит S.Lemma и Flags[i] = 1 и Positions[i] > S.Value.Delta, то: меняем Positions[i] = S.Value.Delta, PositionFlags[i] = S.Value.DeltaFlag.
3. Переход к следующему значению клетки S.

9 Производительность поиска

Эксперименты проведены в соответствии с методикой из [11], но с другими параметрами. Запрос может содержать все виды лемм. Всего 4500 запросов, из них 330 – запросы, которые включают только стоп леммы, 462 – запросы, которые не включают стоп лемм.

Стоп лемм 700, часто используемых лемм 2100, русская и английская морфология.

Проиндексировано 72.5 Гб обычного текста (1 символ – 1 байт). Созданы:

1) Обычный индекс, который для каждой леммы включает все ее словопозиции.

2) Обычный индекс, который для каждой леммы, кроме стоп лемм, включает все ее словопозиции, и дополнительно хранит в составе словопозиции информацию о расположенных близко стоп леммах (*MaxDistance* = 5). Индекс последовательностей стоп лемм (длины последовательностей от 2-х до 5-и), используется для запросов, которые включают только стоп леммы. Расширенные индексы, *ProcessingDistance* от 5 до 7.

Среднее число словопозиций, прочитанных при выполнении запроса: 1) 171 млн. 2) 753 тыс.

Результаты показывают: число прочитанных словопозиций при использовании дополнительных индексов снижено существенно.

10 Структура индекса

Существует два варианта организации инвертированного файла.

1. Использование внешней сортировки слиянием. Вначале тексты читаются, формируются словопозиции и записываются в файл инвертированного индекса в порядке их встречаемости в текстах. Затем, с помощью внешней сортировки файл индекса сортируется таким образом, чтобы словопозиции одной леммы располагались подряд.
2. Легко обновляемые индексы. Словопозиции одной леммы хранятся в наборе блоков. При появлении новых словопозиций этой леммы, идет запись в этот набор блоков, при необходимости добавляются дополнительные блоки. Это позволяет избежать внешней сортировки. См., например, [6].

При реализации дополнительных индексов мы используем легко обновляемые индексы. Детали реализации описаны в [13, 14].

Особенность дополнительных индексов для часто используемых слов заключается в том, что объем данных в индексе, то есть, число словопозиций, для конкретного ключа существенно меньше, чем в обычных индексах.

Это приводит к задаче оптимизации построения индекса, для чего разрабатывается новая подсистема ввода-вывода.

Заметим, что теоретически, для создания дополнительных индексов можно использовать и обычные инвертированные индексы, с использованием внешней сортировки. Но этот вопрос выходит за рамки текущей работы.

11 Новая подсистема ввода-вывода

11.1 Логические файлы

Определяем файл, как объект, поддерживающий операции Запись(*O*, *S*, *B*) и Чтение(*O*, *S*, *B*), где *O* – смещение (адрес) относительно начала файла для записи или чтения, *S* – размер блока данных в байтах, *B* – данные.

На основании файлов операционной системы мы можем создавать логические файлы. Мы имеем дерево, в котором узел – это файл, а его потомки – файлы, от которых он зависит, назовем их файлами-компонентами или дочерними файлами. При записи в файл запись осуществляется в файлы более низкого уровня. Рассмотрим модель подсистемы ввода-вывода, которая состоит из четырех уровней.

Приложение
Распределенные файлы и другие виды логических файлов
Модульный маршрутизатор
Операционная система

Уровень приложения – в нашем случае это компонент системы, который осуществляет запись индекса.

Распределенный файл это особый вид логического файла, предназначенный для оптимизации записи в индекс.

Модульный маршрутизатор, это компонент, который при открытии файла по его имени или другим параметрам, определяет, какой вид логического файла нужно создать.

Какие-то части приложения могут напрямую осуществлять запись в файлы операционной системы, а другие части могут использовать распределенные файлы.

11.2. Распределенные файлы

Распределенный файл формируется на базе нескольких файлов. У каждого файла есть поле *TotalSize* – размер файла.

Распределенный файл представляет собой последовательность блоков разного размера. Размер блоков определяется приложением. Конкретный блок храниться в одном из дочерних файлов.

В составе распределенного файла существует таблица. Ее предназначение, преобразование адреса распределенного файла в адрес одного из дочерних файлов. Таблицу реализуем в виде АВЛ дерева [9]. Ключ в таблице – адрес в распределенном файле, значение – запись, определяющая блок данных, назовем ее описателем блока. Поля описателя блока:

1) *Start* – адрес в распределенном файле.
2) *Size* – размер блока.
3) *Position* – адрес в файле-компоненте.
4) *Index* – номер файла-компонента, в котором хранится блок, определяет файл-компонент.
5) *Next*, *Prev* – переменные для организации циклического списка.

Мы предусматриваем 3 вида файлов-компонентов.

1) Файл для малых операций ввода-вывода, ФМО.
2) Файл для обычных операций ввода-вывода, ФОО.
3) Файл метаданных, ФМ.

Блоки данных сохраняются в ФМО или ФОО. Файл-компонент определяется на этапе первой записи конкретного блока, т. е. записи, которая осуществляется в конец распределенного файла.

АВЛ дерево поддерживает операцию *LowerBound*(*V*), которая возвращает описатель *R*, ключ которого *R.Start* <= *V* ("<=" – меньше или равно), и при этом не существует записи *R2*, *R2.Start* <= *V* и *R2.Start* > *R.Start*.

11.3 Процедура записи в распределенный файл

Осуществляем запись блока (*O*, *S*, *B*) в распределенный файл *D*.

Вначале мы определяем, требуется ли записать блок в конец файла, т. е. будет ли это новый блок. В случае, если часть блока должна обновить существующие данные, а часть выходит за пределы существующих данных, блок делится на два блока по границе существующих данных.

Т. е. если $O = D.TotalSize$, значит это новый блок. Иначе, по крайней мере, часть блока, должна обновить существующие данные.

Если запись осуществляется в существующие данные ($O < D.TotalSize$ и $(O + S) <= D.TotalSize$), то по таблице определяем описатели блоков и запись осуществляется в соответствующие им файлы-компоненты.

Если $O < D.TotalSize$ и $(O + S) > D.TotalSize$, то нужно разбить блок на 2 блока, для каждого из которых применяется один из предыдущих случаев.

Если запись осуществляется в конец файла, то мы по размеру блока определяем целевой файл-компонент. Далее мы создаем для блока описатель в таблице и сохраняем данные в конец файла-компонента.

Для файла обычных операций ввода-вывода мы не предусматриваем никакой особой логики.

11.4 Файл для малых операций ввода-вывода

Файл для малых операций разделяется на плоскости. Плоскость – это блок большого размера, например, 8-32 Мб. Размер плоскости фиксирован. Поля записи *Next* и *Prev* описателя предназначены для организации циклического списка. В этом циклическом списке хранятся описатели, соответствующие одной плоскости.

Будем использовать идеи из [5].

Выделим буфер оперативной памяти, размер которого совпадает с размером плоскости. Данный буфер соответствует активной плоскости.

Как мы увидим далее, плоскости могут освобождаться, данные их них перемещаться в другое место. Организован список свободных плоскостей.

При инициализации буфера активной плоскости активная плоскость определяется путем извлечения из списка свободных плоскостей элемента. Если список пуст, то адрес активной плоскости определяется адресом конца файла, то есть мы добавляем в конец файла новую плоскость.

Если мы записываем в файл новый блок, то записываем его в буфер активной плоскости. Записи блоков в буфер осуществляются последовательно.

Если в буфере активной плоскости недостаточно места для записи нового блока, то мы записываем активную плоскость на диск, по ее адресу, и осуществляем повторную инициализацию буфера и выбор новой активной плоскости.

Если мы должны обновить существующий блок, описатель *R*, то возможны два варианта:

1. $S = R.Size$. В этом случае мы осуществляем запись данных в буфер активной плоскости, обновляем в описателе поле *Position*. Описатель теперь соответствует новой плоскости. Удаляем его из циклического списка старой плоскости и помещаем в циклический список новой плоскости.
2. S не равно $R.Size$. В этом случае запись осуществляем в файл-компонент по адресу $R.Position + (O – R.Start)$. То есть оптимизация записи не применяется в случае частичного обновления блока.

При записи новых данных запись будет преимущественно осуществляться в конец файла.

11.5 Ограничения на уровне приложения

Описанная логика предназначена для оптимизации осуществления малых операций записи. При этом на уровень приложения накладываются ограничения:

1. Уровень приложения должен оперировать блоками данных. Блок данных должен иметь логический смысл на уровне приложения. Приложение должно осуществлять атомарную запись блоков. То есть, если есть какой-то блок данных на уровне приложения, то запись его должна осуществляться за одну операцию ввода вывода. Не должно быть такого, что вначале мы пишем часть, например, половину блока, а затем остальную часть.

   Это достаточно серьезное ограничение, так как приложение в обычном случае, если пишет данные блока за несколько операций ввода вывода, но подряд, то это не несет каких-либо негативных последствий. В нашем же случае, если запись осуществляется в распределенный файл, то части блока могут попасть в разные плоскости. Блок будет фрагментирован, что далее, при чтении приведет к падению производительности.
2. Если приложение ранее осуществляло операцию (O, S, B), то повторные операции должны обновлять данный блок также целиком, то есть, если есть повторная операция записи ($O1$, $S1$, $B1$), где $O1 >= O$, $O1 < (O+S)$, то $O1 = O$, $S1 = S$. Это правило может нарушаться, но в случае нарушения оптимизации записи не будет.

11.6 Процесс контроля заполнения плоскости

При записи (O, S, B) в ФМО, при обновлении существующих данных мы можем теперь осуществлять запись данных в конец активной плоскости. Получается, в той плоскости, где ранее были данные блока, появляется неиспользуемое пространство.

Организуем процедуру контроля заполнения плоскостей. В текущей реализации мы определяем переменную, в которой сохраняется общий объем записанных данных. Когда значение переменной

преодолевает заданный порог, запускается процедура контроля свободного места. Перебираем все плоскости, анализируем, насколько они заполнены. Объем плоскости фиксирован. Объем используемых данных определяется суммированием поля *Size* описателей этой плоскости, что можно сделать, так все описатели плоскости в одном циклическом списке. Если плоскость заполнена менее, чем наполовину (задаем параметром необходимый уровень наполненности плоскости), то переносим все данные ее заполненных блоков в активную плоскость. Текущую плоскость объявляем свободной. Описатели перенесенных блоков обновляются.

12 Преимущества новой подсистемы ввода-вывода

12.1 Замена случайного ввода вывода последовательным

Позволяет вместо случайного ввода-вывода, состоящего из малых операций записи осуществлять, в определенных случаях, последовательный ввод вывод.

Недостаток заключается в наличии процедуры контроля свободного места. За счет чего могут возникнуть дополнительные операции ввода-вывода, однако, это тоже будет последовательный ввод-вывод. За счет замены случайного ввода-вывода на последовательный ввод-вывод улучшается производительность.

Операции записи большими блоками и операции чтения не оптимизируются.

12.2 Оптимизация для SSD.

Для SSD массивный случайный ввод-вывод, состоящий из большого числа малых операций записи, может негативно влиять на производительность. Именно такой вид ввода-вывода мы заменяем последовательным.

12.3 Возможность организации многоуровневого хранения для создания индекса

Метод предусматривает возможность размещать файлы для разных видов операций на разных носителях. Например, случайный ввод-вывод можно перенести на SSD, с учетом пункта 12.2.

Заметим, что в настоящее время в системах хранения предусматривается автоматическое многоуровневое хранение данных. Данные, к которым обращаются чаще, перемещаются на более быстрые носители. Однако система хранения не знает об особенностях приложения и за счет ручного управления хранением данных можно повысить производительность.

12.4 Сохранение производительности поиска

При условии выполнения ограничений, накладываемых на уровень приложения, производительность поиска не ухудшится.

12.5 Распараллеливание ввода-вывода

На построенной модели можно организовать хранение индекса на нескольких носителях и распараллеливание ввода-вывода на них.

13 Применение новой модели ввода вывода к построению индекса

13.1 Организация инвертированного индекса

Как было сказано ранее, мы используем способ создания индекса, описанный в [13, 14].

Инвертированный индекс представляет собой файл, который разбит на блоки фиксированного размера – кластеры. Цепочка кластеров или поток – это набор кластеров, в котором хранятся словопозиции. В простейшем случае словопозиции одной леммы хранятся в одном потоке.

Базовая идея заключается в том, что мы создаем один кластер для словопозиций леммы и записываем их в него последовательно. Когда свободное место в кластере заканчивается, создается новый кластер, в текущем кластере прописывается ссылка (номер кластера) на новый кластер.

Необходимо, чтобы блоки одной леммы располагались преимущественно подряд для снижения числа операций ввода-вывода при поиске.

Поэтому, введено понятие блока кластеров, это кластеры, которые идут в файле подряд. Зафиксируем максимальный размер блока кластеров, например, 16 Мб. Рассмотрим блоки кластеров, размером в 1, 2, 4, 8 и т.д. кластеров.

Вначале используется блок, состоящий из одного кластера. Когда свободное место в нем заканчивается, выделяется блок, размером, в два раза больше. Данные из существующего кластера переносятся в половину нового блока. Так делается, до достижения максимального размера блока. После чего мы выделяем уже новый блок, и поток состоит уже из двух блоков. Также есть возможность использования части кластера для одного потока, а другой части для другого.

Нужно, чтобы первоначальная запись блока кластеров осуществлялась атомарно, за одну операцию ввода вывода. Чтобы в распределенном файле был правильно определен целевой файл-компонент и для блока кластеров был создан один описатель.

13.2 Кеш инвертированного индекса

Считаем, что одновременно в памяти мы храним не менее одного кластера для одной леммы. В качестве одного из вариантов выполнения данного условия можно использовать кэш, размер кэша должен быть не меньше, чем число лемм, обрабатываемых на текущий момент, умноженное на размер кластера. Организуем таблицу, ключ в которой – это номер кластера. Значение – структура, содержащая адрес буфера, размером в один кластер,

с данными, и дополнительные поля, например, признак того, что данные изменены.

Поскольку в кэше оперируем единичными кластерами, нужно обеспечить запись данных из кэша в файл таким образом, чтобы кластеры одного блока кластеров сохранялись, например, при выгрузке из кэша, за одну операцию ввода-вывода.

Для этого, в структуру, описывающую кластер в кэше, добавим поле – номер операции ввода вывода.

Введем переменную – счетчик, которая будет увеличиваться на 1 при каждой последующей операции ввода-вывода. Когда осуществляется первоначальная запись блока кластеров, она осуществляется за одну операцию ввода вывода и значение номера операции ввода вывода одинаковое для всех кластеров блока в таблице кэша.

13.3 Учет соединения записей (write coalescing) на уровне распределенного файла

При сбросе кэша на диск, записи отдельных кластеров кэша могут объединяться в одну для оптимизации процесса записи. При этом кластеры могут соответствовать блокам разного вида (ФМО или ФОО). Это нужно учитывать при организации распределенного файла. Если осуществляется запись буфера *B* в распределенный файл и при этом должны быть обновлены несколько блоков распределенного файла, то нужно вначале разделить буфер *B* на части, каждая из которых соответствует одному блоку распределенного файла. Затем для блоков одного вида осуществить обратное соединение записей, на основании частей буфера *B*, которые соответствуют этим блокам, если это возможно.

14 Заключение

Рассмотрены методы создания дополнительных индексов для поиска фраз, включающих часто используемых слова. Эксперименты показывают, что число словопозиций, которые должны быть обработаны при выполнении поискового запроса, может быть снижено более чем в 10 раз за счет применения дополнительных индексов. Дан новый алгоритм поиска для случая запроса, включающего хотя бы одно часто используемое слово, но без стоп слов. Представлена новая подсистема ввода-вывода для оптимизации создания таких индексов.

Литература

[1] Bahle D., Williams H.E., Zobel J.; Efficient Phrase Querying with an Auxiliary Index. In Proc. ACM-SIGIR Conf. on Research and Development in Inform. Retrieval, Finland, 2002, p. 215–221.

[2] Chang M., Chung Keung Poon. Efficient Phrase Querying with Common Phrase Index. ECIR 2006, LNCS 3936, Springer-Verlag Berlin Heidelberg, 2006, p. 61–71.

[3] Shashank Gugnani, Rajendra Kumar Roul. Triple Indexing: An Efficient Technique for Fast Phrase Query Evaluation. International Journal of Computer Applications. Vol 87, No. 13, 2014.

[4] Prywes N. S., Gray H. J; The organization of a Multilist-type associative memory, IEEE Trans. on Communication and Electr., 1963, 68, p. 488–492.

[5] M. Rosenblum and J. Ousterhout. The Design and Implementation of a Log-Structured File System. ACM Trans. Comp. Syst., 10(1), 1992, p. 26–52.

[6] Tomasic A., Garcia-Molina H., Shoens K.; Incremental updates of inverted lists for text document retrieval, In Proc. ACMSIGMOD Int. Conf., Minnesota, 1994, p. 289–300.

[7] Williams H. E., Zobel J., Bahle D.; Fast phrase querying with combined indexes. ACM TOIS. Vol 22, No. 4, 2004, p. 573–594.

[8] George Kingsley Zipf. Relative frequency as a determinant of phonetic change. Harvard Studies in Classical Philology, Vol 40, 1929, p. 1–95.

[9] Адельсон-Вельский Г. М., Ландис Е. М. (1962). Один алгоритм организации информации. Докл. АН. СССР, 146, с. 263-266.

[10] Веретенников А.Б. О поиске фраз и наборов слов в полнотекстовом индексе, Системы управления и информационные технологии, №2.1(48), 2012, с. 125–130.

[11] Веретенников А. Б. Использование дополнительных индексов для более быстрого полнотекстового поиска фраз, включающих часто встречающиеся слова, Системы управления и информационные технологии, №2(52), 2013, с. 61–66.

[12] Веретенников А. Б. Создание дополнительных индексов для более быстрого полнотекстового поиска фраз, включающих часто встречающиеся слова. Системы управления и информационные технологии, №1(63), 2016, с. 27–33.

[13] Веретенников А. Б. Создание легко обновляемых текстовых индексов, RCDL'2008. Дубна: ОИЯИ, 2008, с. 149–154.

[14] Веретенников А. Б. Эффективная индексация текстовых документов с использованием CLB-деревьев, Системы управления и информационные технологии, №1.1(35), 2009, с. 134–139.

Using additional indexes of frequently used words for full-text search

Alexander B. Veretennikov

Different words can occur in texts with different frequency. We describe additional indexes, intended for speeding up search in case the search query contains frequently used words. We defined several groups of words and developed different methods for each group. Index writing optimization is given. For the case when search query is a set of frequently used words, a search algorithm is explained.

Towards Text Processing System for Emergency Event Detection in the Arctic Zone

© Dmitriy Deviatkin © Artem Shelmanov

Federal Research Center "Computer Science and Control" of
Russian Academy of Sciences,
Moscow, Russia

devyatkin@isa.ru shelmanov@isa.ru

Abstract

We present the ongoing work on text processing system for detection and analysis of events related to emergencies in the Arctic zone. The peculiarity of the task consists in data sparseness and scarceness of tools / language resources for processing such specific texts. The system performs focused crawling of documents related to emergencies in the Arctic region, text parsing including named entity recognition and geotagging, and indexing texts with their metadata for faceted search. The system aims at processing both English and Russian text messages and documents. We report the preliminary results of the experimental evaluation of the system components on Twitter data.

Keywords: focused crawling, event detection, monitoring, named entity recognition, text processing, information search

1 Introduction

Due to ever-growing amounts of data available on the web, monitoring and searching in textual streams is still one of the most urgent problems today that has inspired researchers to develop many general-purpose information-retrieval methods and systems. However, the development of applications for specific domains often reveals lack of suitable techniques that could address challenging tasks arising in these domains, which require significant research.

This paper describes an ongoing development of a search and monitoring system for a specific domain and a task. It is oriented on detection and analysis of emergency events in the Arctic zone. Since a lot of textual information is generated during emergencies and crises, as during major events of other types, it is crucial to have automated tools for filtering and processing of unstructured textual data for support of search and rescue operations, as well as for helping people in affected areas. The Arctic zone is a hard but important and promising region that has a lot of potential for the development. The remarkable peculiarity of the chosen domain is data sparseness and scarceness of tools / language resources for processing such specific data, which poses a difficult problem.

The most significant features of the system are focused crawling and faceted search.

Since it is impossible to store all available data on the web, the developed system is designed to accumulate only data related to emergencies in the Arctic zone from multiple textual streams. The sources of such information include but are not limited to mass media, social networks, reports (e.g., official sources like national transportation safety boards[1,2]). The focused crawler is intended to narrow down the amount of indexed text and extract basic metadata of downloaded documents. At first sight, the problem of crawling messages about emergency events is very similar to topic crawling. The key difference lies in the fact that emergency related messages can be devoted to multiple topics and the composition of these topics can change over time. It means that using the ordinal topical approaches leads to inappropriate accuracy and laboriousness of the crawling process. To mitigate this problem, we have implemented the following ideas in the proposed framework:

- Multiple topic crawlers with narrow focuses outperform a single data collecting process in terms of recall.
- Geographical coordinate extracting and considering them for further filtering improve the accuracy of the crawling process. One could get topically irrelevant, but important messages from emergency zone.
- Topic models for crawled texts could be periodically built and verified for better tracking of topic shifts in text streams.
- Reposts and fuzzy duplicates can be effectively detected via inverted full-text indices [28].

The faceted search provides the abilities to retrieve and analyze texts in different perspectives: topic, time,

Proceedings of the XVIII International Conference «Data Analytics and Management in Data Intensive Domains» (DAMDID/RCDL'2016), Ershovo, Russia, October 11 - 14, 2016

[1] http://www.tsb.gc.ca/eng/rapports-reports/marine/index.asp

[2] http://www.ntsb.gov/investigations/AccidentReports/Pages/marine.aspx

location, relations with the given object, etc. The developed system performs deep natural language processing of texts (including syntax parsing and semantic role labeling), named entity recognition, as well as geotagging. The extracted metadata is indexed for the faceted search.

We evaluated the developed subsystems for geotagging, crawling, and faceted search on the data acquired from Twitter. Although this social network accumulates only short messages and is not designed for providing data for the considered tasks, many researchers, as shown in Section 2, demonstrated that tweets could be a useful source of information about emergencies. When common communication services are down, Twitter provides a channel, which is used by affected people and emergency response teams [22]. Therefore, we used messages crawled from Twitter for preliminary experiments, testing our approaches, and evaluation of the system components. However, we note that the developed system is designed to handle all sorts of textual information, not just short messages.

The rest of the paper is organized as follows. Section 2 reviews the related work about monitoring emergency events with help of social networks and focused crawling. Section 3 describes the details of the system in development; it presents the natural language processing pipeline, method for focused crawling, and faceted search techniques. In section 4, the results of the preliminary experiments are presented and discussed. Section 5 concludes and outlines the future work.

2 Related work

The problem of event detection in text streams has a lot of attention from the research community. Methods that were developed to address this problem were applied to many domains. One of them is monitoring emergencies. It was noticed that mass emergencies initiate the intensive exchange of information in social networks. This immense text stream contains cues about a situation in an affected area, infrastructure damage, human casualties, requests and proposals for help. It is a crucial information that can enhance the situation awareness [18] of both affected people and participants of rescue operations. However, it is mixed up with heavy noise: irrelevant or useless messages. Therefore, to put it to good use, new methods and technologies are required. The need of such technologies became apparent, which facilitated the development of many diverge systems for mining emergency related information in social networks. We review the most significant recent work on such systems.

Papers [20] and [17] present an information flow monitoring system Twittris designed for processing of short messages from mass and social media, as well as SMS-messages. Researchers tested the system on Twitter data. The system crawls messages from Twitter using a set of keywords, which is expanded over time by the most significant n-grams extracted from acquired messages. The system extracts the spatial and temporal information, as well as topics, which are used for message clustering. The clusters are considered as events found in an information stream. Researchers tested the system on the data acquired during hurricane Sandy. They showed that the system could be used for searching messages from affected people considering their location.

Another monitoring system SensePlace2, described in [12], specializes on analysis of the geographical data extracted from tweets. The system aims at improving the situational awareness during search and rescue operations. The main goal of the system is text stream filtering and searching of messages related to the given topic, place, and time. The system utilizes the geographical tags, as well as the information extracted from message texts. Besides text, Senseplace2 also indexes geographical and temporal information of messages. This enables the system to filter a message stream by place and time and build analytical reports for topic-time-location data. Senseplace2 can visualize results in different ways: as a common search result list, present them on a time scale as a histogram, and visualize results on a heat-map, which displays the intensity of messages about particular topic near the given location. Researchers tested the system using data related to the Haiti earthquake. They showed that SensePlace2 could be useful for finding refugee streams that are not represented in official sources.

In [24], researchers present a method for classification of messages acquired from a message stream. They demonstrate its capabilities of finding useful emergency related messages on Twitter data. The method can classify messages as useful and non-useful via standard supervised machine learning methods (Naïve Bayes and Maximum entropy). The most remarkable thing is a feature set used for training. Besides low-level features, they also conducted experiments with high-level features like message objectivity, whether it is personal or impersonal, whether it is formal or informal. The authors show that high-level features substantially improve the quality of classification. The out-of-domain evaluation showed accuracy from 30 to 80%. The experiments are conducted on the data acquired during Haiti earthquakes, USA wildfires, and floods.

The system EMERSE (Enhanced Messaging for the Emergency Response Sector) [4] collects messages from different sources, translates them, and classifies them into topics for better search and filtering. EMERSE consists of a smartphone application, a Twitter crawler, a translation subsystem, and a subsystem for classification. The smartphone application is considered to simplify a process of collecting messages and their metadata such as location, time, and associated media files (photo, video). Besides, the system crawls Twitter considering timestamps and eliminating duplicates (reposts). EMERSE classifies messages into multiple classes using support vector machine. In [4], authors experimented with different features and feature selection methods: bag of words, feature abstraction methods [21], Latent Dirichlet Allocation (LDA), and others. The system was tested on a collection of messages submitted to the

Ushahidi[3] web-service during the Haiti earthquake. In this example, the authors demonstrate that EMERSE can improve coordination of people during the emergencies.

In [25, 26], a system ESA (Emergency Situation Awareness) is presented. It can monitor social networks and blogs in real time and visualize information about different emergencies. The main task of the system is to enhance situational awareness of people in an affected area. The system is oriented on New Zealand and Australia regions. ESA gathers tweets and detects topical bursts in information streams. The retrospective data is used for building a language model, which is applied for the further burst detection. The algorithm searches lexis that has a very diverge distribution comparing to the language model. For convenient representation of bursts for end-users, ESA performs thematic clustering of messages. The system also selects informative messages that signal about emergencies, destructions, and requests for help. ESA has a component that extracts relevant spatial data using explicit geotags of messages (GPS-coordinates received from a smartphone) and implicit information, found in user profiles. The conversion from geographical names to coordinates is performed by Yahoo geo-service[4] (retired today). ESA also performs named entity recognition: it extracts names of organizations, names of people, geographical entities, dates, and timestamps. All these data can be visualized on a map, which could be useful for providing better representation of found events for end users. Visualization of data in ESA is also enhanced with media files (images, videos), extracted from messages. The authors tested ESA in Australian crisis center, which is responsible for monitoring of natural disasters and other national security threats.

AIDR[5] (Artificial Intelligence for Disaster Response) is an open-source platform for classification of messages related to emergencies [9]. The system detects messages about different topics: infrastructure damage, casualties, required or available donations. The authors point out that classifiers trained on the data collected during one disaster perform badly on the data acquired from new disasters. They address this problem by introducing human annotation into the process of adapting the system to new tasks. When a new emergency happens, the system should be retrained. The training dataset for supervised machine learning is composed from the old labeled data and the data urgently annotated via crowdsourcing services. The systems have elements of active learning; it chooses for human annotation the most informative samples that can significantly leverage classification performance. The authors tested the system on the collection of messages related to Pakistan earthquake in 2013.

TEDAS [10] is the system for emergency detection via focused crawling of Twitter messages. TEDAS collects topic relevant messages using the Twitter search API. The system uses the original crawling strategy that consists in dynamic shifting of crawler focus.

Another system for vertical search of information about emergencies is described in [27]. The system includes a focused ontology-based crawler. An extensive ontology describing various emergencies is designed for the crawler.

It is also worth mentioning Tweedr [2] – an open-source system that can find informative messages from Twitter for information support of people involved in rescue operations. It can distinguish general messages from the ones that have particular information about infrastructure damage and human casualties. Another recent effort in constructing tweet classification system is described in [5]. The authors use deep natural language processing techniques and rich set of features to determine whether a message contains information about damage dealt during natural disasters. In [14], an approach for construction of crisis-related terms is proposed. Authors used pseudo-relevance feedback mechanisms to expand a number of seeding terms during crawling, which results in recall improvement of retrieving of messages related to mass emergencies. Another lexicon called EMTerms is described in [23]. The authors claim that it is the biggest crisis-related lexicon for Twitter analysis so far.

Solutions for monitoring events in text streams heavily depend on focused crawling techniques. We review some of the state-of-the art approaches below.

ICrawl system [7] is a framework for focused crawling of social networks. It adopts ontology based crawling strategy. The novel feature of this system is a usage of Internet search engines for generation of bootstrap crawling points. In [3], researchers propose a distributed crawler for continuous message-gathering from particular user communities, which can circumvent limits of Twitter API. In [11], automatic Topic-focused Monitor is presented. It samples tweets from the message stream and selects keywords to track target topics based on the samples.

The review shows that there are a plenty of systems for monitoring emergency related events in textual streams intended to improve situational awareness of affected people and rescue teams. In our work, we consider a particular geographical region – the Arctic zone, which complicates focused crawling and filtering of data. Many aforementioned systems specialize on narrow problems like message classification, whereas our research is oriented on the development of a full-stack system that solves many tasks: from focused crawling and information extraction, to faceted search leveraged with spatial and temporal metadata. Unlike the aforementioned systems, the framework proposed in this paper is oriented on processing messages in both English and Russian languages. This is significant because of the large area of the Arctic territories of Russia. We note that many systems use Twitter data for evaluation, and we also use this approach in our work.

[3] https://www.ushahidi.com/

[4] https://developer.yahoo.com/boss/geo/

[5] http://aidr.qcri.org/

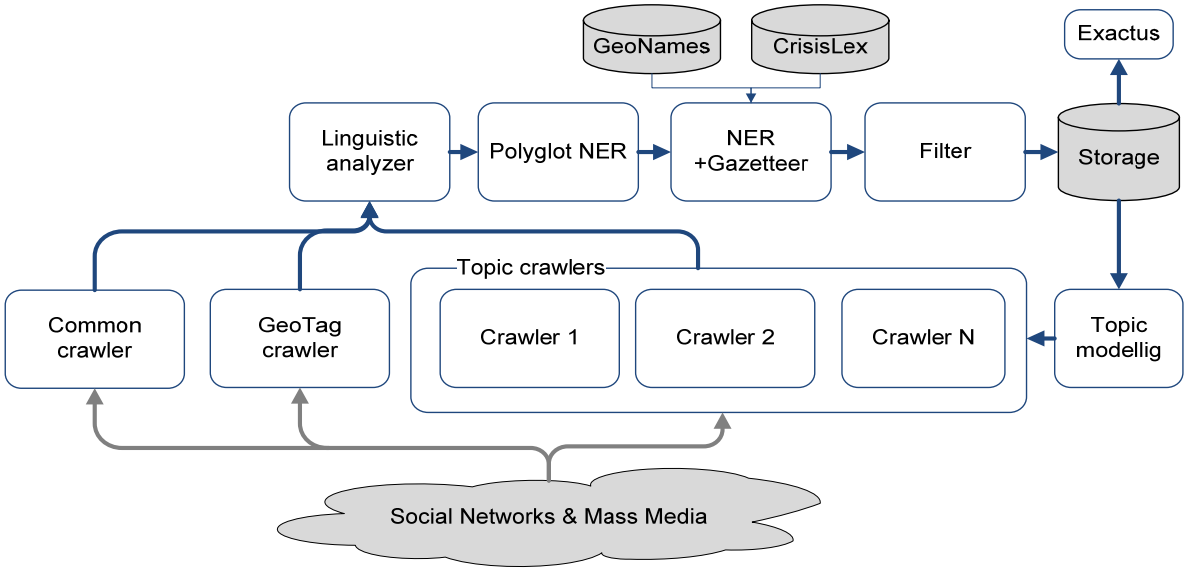


Figure 1 Framework for crawling of emergency messages

3 System components

3.1 Natural language processing pipeline

The system performs deep natural language processing of Russian and English texts. Besides basic processing tools, the pipeline also includes syntax parsing, semantic role labelling, and named entity recognition.

The basic analysis for Russian texts is performed by AOT.ru[6]. This framework is used for tokenization, sentence boundary detection, POS tagging, and lemmatization, including morphological disambiguation. We use MaltParser[7] trained on SynTagRus [13] for dependency parsing of Russian texts and our semantic parser for semantic role labelling [19]. The same types of linguistic analysis of English texts are performed via Freeling [16]. Note that the syntax and semantic annotations are used for information search (see section 3.3).

For the basic named entity recognition, we used Polyglot NER framework [1]. It implements language agnostic approach and due to this provides named entity recognition for many languages including English and Russian. It produces annotations for locations, organizations, and person names. However, we found that the basic NER processor is not suitable for extracting toponyms related to a particular region (e.g., Arctic zone); it yields low recall in this task. Therefore, we complemented Polyglot with a gazetteer.

The gazetteer was created on the basis of Geonames[8] database. It contains more than 11 million geographical locations of different types around the world with their names (in many languages including Russian and English), geographical coordinates, and other metadata. From Geonames, we extracted location names that are situated on the north of the 60th latitude. The gazetteer uses these data to mark spatial information in texts. It also implements rather simple rules to filter out common false positives that take into account parts of speech and capitalization of words.

We also tag crisis-related lexis in texts; it enhances and simplifies filtering and search. The data for this purpose is taken from CrisisLex lexicon, proposed in [14].

3.2 Focused crawling framework

We deal with several social networks, such as Twitter, Facebook, and VKontakte, and with some news feeds (ArcticInfo, BarentsObserver, BBC, etc.) These sources provide different kinds of content. The Twitter provides API for crawling of recent messages by keywords. However, the limitations of the API make the topical crawling process challenging. Since results commonly contain much irrelevant noise, additional filtering is necessary. We access Facebook and VKontakte primarily via links in twitter messages that are considered topically relevant. The news feeds have a static structure, therefore, they can be processed by a common crawler with a preliminary created static task. The data acquired from news feeds do not need topical filtering, because the crawling task can be restricted to process only relevant sections. Since we deal with a number of heterogeneous sources, we use several kinds of crawlers (see Fig. 1).

The first type is a GeoTag crawler. It is used for collecting messages from Twitter with specified coordinates. Tweets may include geographical coordinates or geo-tags, which could be used for localization of their authors. We filter all messages, whose geo-tag latitude is less than 60 degrees.

The second type is a Topic crawler. These crawlers download topically relevant messages from Twitter with unspecified coordinates. Each topic crawler has lists of "permissive" and "restrictive" terms that are fed to

[6] http://aot.ru/
[7] http://maltparser.org/
[8] http://www.geonames.org/

Table 1 Examples of topics for crawled data

No	Keywords	Relevant
1	Bay, charity, Amazon, Antarctica, cdnpoli.	False
2	Starling, Tuktoyaktuk, community, visit, bird, southern, blackbird.	True
3	Ice, national, ship, circle, arctics, photography, day, june, pewenvironment.	True
4	Rescue, buntings, air, guardsmen, squadron, cranes, divers, spot.	True
5	Haha, dart, Trump, meepismurder, white, sales, gauges, street.	False
6	Icebreaker, Nunavut, hardy, apithanny, piece, fascinating, blue, warming, bear.	True
7	Home, conservation, thebigbidtheory, may, island, science, hydrazine.	False
8	Spring, noaa, climatechange, water, super, sail, challenge, Mediterranean.	False
9	Arctic, Alaska, skuas, Greenland, road, amb, melt, Anchorage, Bering.	True
10	Life, natgeomag, trip, journey, remote, team, chukchi, collaborating.	True

Twitter search API. In the initial steps, several bootstrap terms are used for defining a target topic. The challenge lies in limitations of topic search API, provided by Twitter. It restricts a size of a query and a response, which leads to insufficient recall of the crawling process. The simplicity of the query language causes the low precision and recall of the collected data. We use multiple topic crawlers with different keyword subsets to solve the insufficient recall problem. NER and filtering are used to improve the precision.

The last type of crawlers is a common crawler. It collects data from topically related sections of news feeds. The crawlers of this type can also download pages from VKontakte and Facebook referenced by relevant Twitter posts.

The whole schema of data processing in our framework is the following. In the first step, messages are collected by GeoTag and Topic crawlers. In the second step, we apply linguistic analyser, NER, and gazetteer to the collected texts. Then, we filter all messages that do not contain any crisis lexis, toponyms, and geotags. URLs from the remaining messages are fed to the common crawler that also processes topically related news feeds. The selected useful messages and documents are indexed by the Exactus search engine [15].

For Topic crawler, we build a topic model [8] of the crawled messages every several days. It helps to track topic shifts in the message stream. We summarize topic content with a keyword cloud and a set of the most significant messages from the cluster. Then each topic is marked as relevant or irrelevant by several assessors (see Table 1). We define the following types of posts as relevant:

1. Posts about arbitrary events (past, current and planned) and locations in the Arctic.
2. Arbitrary posts from users, who currently are in the Arctic zone.

The most significant terms from the relevant topics are sent to “permissive” keyword collections of topic crawlers, and terms from irrelevant topics are sent to “restrictive” ones. Thus, the crawling process becomes responsible to trend shifts.

3.3 Faceted search

The faceted search became a backbone for professional search applications [6]. In this type of search, users can iteratively specify queries using metadata and keywords extracted from search results of previous iterations. Additionally, search results could be filtered using different sets of meta fields that can be static or dynamic.

In the developed system, the faceted search is powered by the Exactus technology [15]. Its main advantage lies in ability to efficiently index rich linguistic information including syntax relation, semantic roles, or other types of semantic annotations extracted from natural language text (e.g. named entities). This enables phrase search (results have to contain given syntactically connected phrases) semantic search (results are ranked taking into account semantic similarity of the query and indexed documents). We take advantage of this technology by introducing indexing by geographical tags, timestamps, and emergency-related tags. This provides the ability to filter results efficiently by semantic information like location, time, organizations, persons, and topics. It also provides the ability to retrieve information with certain tags filtered by other metadata producing the results that can be sifted with consequent queries.

4 Evaluation of system components

We have conducted a series of experiments to assess the quality of the created components for focused crawling, named entity recognition, and faceted search. The source of the data for evaluation is Twitter social network. The experimental dataset contains approximately 100 thousand messages in English and Russian. In the first experiment, we assessed accuracy of the proposed focused crawling framework. More specifically, we evaluated the quality of filtering. We labelled several subsets of posts devoted to accidents in Alaska and Bering Sea. Each post from the subsets was labelled by three assessors to reach sufficient coherence of the test data. We have not applied a cross-validation approach here because the labelling was not used for the crawler training, just for testing. The standard measures for supervised learning: precision, recall and F_1-score, were used for each subset. Macro-averaging was used to evaluate the result assessments. Table 2 refers to the results of the crawling without and with filtering as “Impure data” and “Filtered data” correspondingly.

Applying the proposed filtering technique results in the substantial growth of the precision without the

significant decrease of the recall. This means that during the crawling process we do not lose much topically relevant data but substantially decrease the stored noise. We decided to choose a fairly soft filtering because, although a stricter procedure would improve the precision, it would also imply a more significant recall drop, which contradicts the purpose of the monitoring system.

Table 2 Focused crawling evaluation

	P	R	F_1
Impure data	0.26	1.00	0.41
Filtered data	0.57	0.94	0.70

In the second experiment, we estimated the performance of named entity recognition performed by Polyglot and gazetteer. We labelled all location mentions in 300 tweets that were downloaded by the Topic crawler and measured precision, recall, and F_1-score for extraction of spatial entities (Table 3).

Table 3 NER evaluation (on locations)

	P	R	F_1
Polyglot	0.78	0.57	0.66
Gazetteer	0.78	0.74	0.76
Polyg.+gazetter	0.76	0.82	0.79

Results show that proposed Gazetteer significantly outperforms Polyglot on location extraction in terms of recall. The knowledge source of Polyglot is Wikipedia that does not have the full coverage of locations. We conclude that it is reasonable to use the gazetteer and Polyglot together for the maximum performance.

In the last experiment, we assessed the performance gain of the information search achieved by using the proposed emergency faceted search method in comparison to the baseline algorithm. We deployed the Exactus full-text search algorithm without filtering by tag locations as the baseline. For the evaluation, we applied the NDCG score and peer reviewing approach. The results are presented in Table 4.

Table 4 Faceted search evaluation

	3-DCG	5-DCG	10-DCG
Faceted	0.76	0.76	0.70
Baseline	0.61	0.55	0.53

It was revealed that use of location and crisis tags for faceted search significantly improves the quality of ranking when searching posts about emergencies.

5 Conclusion

We presented an automated framework for crawling and processing textual documents about emergency events in the Arctic zone. The main functions of the proposed framework are focused crawling and faceted search that takes into account information about geographical locations and timestamps of messages. With the data crawled from Twitter, we experimentally demonstrated that the framework provides the basic abilities for analysis of message streams about emergencies in the restricted area.

In the future work, we are going to incorporate into the natural language processing pipeline components that extract information about ships and planes in the Arctic zone. Bulk information is available openly on the web (e.g., MarineTraffic service[9]). Tagging ship names and their coordinates in document and message streams potentially can improve the quality of emergency event detection and enhance the situation awareness.

We are going to accumulate more retrospective data from social networks and other sources to increase the recall of the crawling process. Among many other types of information sources, collections of reports from rescue services are the most prospect supplement for the crawling. Another way to improve topic crawling is detection of users and groups in social networks that constantly post topically relevant messages. This could be done semi-automatically by building topical models on users and groups. We are also going to create visualization tools for geotagged messages that can present events on the map.

Acknowledgments

The project is supported by the Russian Foundation for Basic Research, project number: 15-29-06045 “ofi_m”.

References

[1] Rami Al-Rfou, Vivek Kulkarni, Bryan Perozzi, and Steven Skiena. Polyglot-NER: Massive multilingual named entity recognition. In Proceedings of the 2015 SIAM International Conference on Data Mining. SIAM, 2015.

[2] Zahra Ashktorab, Christopher Brown, Manojit Nandi, and Aron Culotta. Tweedr: Mining Twitter to inform disaster response. Proceedings of ISCRAM, pages 354–358, 2014.

[3] Matko Boanjak, Eduardo Oliveira, José Martins, Eduarda Mendes Rodrigues, and Luís Sarmento. Twitterecho: a distributed focused crawler to support open research with twitter data. In Proceedings of the 21st international conference companion on World Wide Web, pages 1233–1240. ACM, 2012.

[4] Cornelia Caragea, Nathan McNeese, Anuj Jaiswal, Greg Traylor, Hyun-Woo Kim, Prasenjit Mitra, Dinghao Wu, Andrea H. Tapia, Lee Giles, Bernard J. Jansen, et al. Classifying text messages for the Haiti earthquake. In Proceedings of ISCRAM, 2011.

[5] Stefano Cresci, Maurizio Tesconi, Andrea Cimino, and Felice Dell’Orletta. A linguistically-driven approach to cross-event damage assessment of

[9] http://www.marinetraffic.com/

natural disasters from social media messages. In Proceedings of the 24th International Conference on World Wide Web Companion, pages 1195–1200. International World Wide Web Conferences Steering Committee, 2015.

[6] Pavlos Fafalios and Yannis Tzitzikas. Exploratory professional search through semantic post-analysis of search results. In Professional Search in the Modern World, pages 166–192. Springer, 2014.

[7] Gerhard Gossen, Elena Demidova, and Thomas Risse. The iCrawl Wizard – supporting interactive focused crawl specification. In Advances in Information Retrieval, pages 797–800. Springer, 2015.

[8] Thomas Hofmann. Probabilistic latent semantic indexing. In Proceedings of the 22nd annual international ACM SIGIR Conference on Research and Development in Information Retrieval, pages 50–57. ACM, 1999.

[9] Muhamma-d Imran, Carlos Castillo, Ji Lucas, Patrick Meier, and Sarah Vieweg. AIDR: Artificial intelligence for disaster response. In Proceedings of the companion publication of the 23rd International Conference on World Wide Web Companion, pages 159–162, 2014.

[10] Rui Li, Kin Hou Lei, Ravi Khadiwala, and Kevin Chen-Chuan Chang. Tedas: A twitter-based event detection and analysis system. In Data engineering (ICDE), 2012 IEEE 28th international conference, pages 1273–1276. IEEE, 2012.

[11] Rui Li, Shengjie Wang, and Kevin Chen-Chuan Chang. Towards social data platform: Automatic topic-focused monitor for twitter stream. Proceedings of the VLDB Endowment, 6(14):1966–1977, 2013.

[12] Alan M. MacEachren, Anuj Jaiswal, Anthony C. Robinson, Scott Pezanowski, Alexander Savelyev, Prasenjit Mitra, Xiao Zhang, and Justine Blanford. Senseplace2: Geotwitter analytics support for situational awareness. In Proceedings of Visual Analytics Science and Technology (VAST) on IEEE Conference, pages 181–190, 2011.

[13] Joakim Nivre, Igor M. Boguslavsky, and Leonid L. Iomdin. Parsing the SynTagRus treebank of Russian. In Proceedings of the 22nd International Conference on Computational Linguistics (Coling 2008), pages 641–648, 2008.

[14] Alexandra Olteanu, Carlos Castillo, Fernando Diaz, and Sarah Vieweg. CrisisLex: A lexicon for collecting and filtering microblogged communications in crises. In Proceedings of ICWSM, 2014.

[15] Gennady Osipov, Ivan Smirnov, Ilya Tikhomirov, Ilya Sochenkov, and Artem Shelmanov. Exactus expert – search and analytical engine for research and development support. In Novel Applications of Intelligent Systems, pages 269–285. Springer, 2016.

[16] Lluís Padró and Evgeny Stanilovsky. Freeling 3.0: Towards wider multilinguality. In Proceedings of the Language Resources and Evaluation Conference (LREC 2012). ELRA, 2012.

[17] Hemant Purohit and Amit P Sheth. Twitris v3: From citizen sensing to analysis, coordination and action. In Proceedings of ICWSM, pages 746–747, 2013.

[18] Nadine B Sarter and David D Woods. Situation awareness: A critical but ill-defined phenomenon. The International Journal of Aviation Psychology, 1(1):45–57, 1991.

[19] A. O. Shelmanov and I. V. Smirnov. Methods for semantic role labeling of Russian texts. In Computational Linguistics and Intellectual Technologies. Papers from the Annual International Conference "Dialogue" (2014), number 13, pages 607–620, 2014.

[20] Amit P. Sheth, Hemant Purohit, Ashutosh Sopan Jadhav, Pavan Kapanipathi, and Lu Chen. Understanding events through analysis of social media. Kno.e.sis Center, Wright State University, Tech. Rep., 2010.

[21] Adrian Silvescu, Cornelia Caragea, and Vasant Honavar. Combining super-structuring and abstraction on sequence classification. In Proceedings of ICDM, pages 986–991. IEEE, 2009.

[22] Juan Sixto, Oscar Pena, Bernhard Klein, and Diego López-de Ipina. Enable tweet-geolocation and don't drive ERTs crazy! Improving situational awareness using Twitter. Proceedings of SMERST, pages 27–31, 2013.

[23] Irina Temnikova, Carlos Castillo, and Sarah Vieweg. Emterms 1.0: a terminological resource for crisis tweets. In ISCRAM 2015 Proceedings of the 12th International Conference on Information Systems for Crisis Response and Management, 2015.

[24] Sudha Verma, Sarah Vieweg, William J Corvey, Leysia Palen, James H Martin, Martha Palmer, Aaron Schram, and Kenneth Mark Anderson. Natural language processing to the rescue? Extracting "situational awareness" tweets during mass emergency. In Proceedings of ICWSM, pages 385–392, 2011.

[25] Jie Yin, Sarvnaz Karimi, Bella Robinson, and Mark Cameron. ESA: emergency situation awareness via microbloggers. In Proceedings of the 21st ACM International Conference on Information and Knowledge Management, pages 2701–2703. ACM, 2012.

[26] Jie Yin, Andrew Lampert, Mark Cameron, Bella Robinson, and Robert Power. Using social media to enhance emergency situation awareness. IEEE Intelligent Systems, (6):52–59, 2012.

[27] Li Zheng, Chao Shen, Liang Tang, Chunqiu Zeng, Tao Li, Steve Luis, Shu-Ching Chen, and Jainendra K Navlakha. Disaster SitRep – a vertical search engine and information analysis tool in disaster management domain. In Information Reuse and Integration (IRI), 2012 IEEE 13th International Conference, pages 457–465. IEEE, 2012.

[28] Denis Zubarev and Ilya Sochenkov. Using sentence similarity measure for plagiarism source retrieval. In CLEF (Working Notes), pages 1027–1034, 2014.

Данные: интеграция, совместное использование, получение от датчиков

Data Integration, Sharing and Sensing

Управление семантическими активами и их повторное использование для решения задач информационного взаимодействия

© Ю. М. Акаткин © Е. Д. Ясиновская © М.Г. Бич © А.В. Шилин

Российский экономический университет им. Г.В. Плеханова
г. Москва
uakatkin@yandex.ru elena@semanticpro.org misha@e-projecting.ru a.shilin@e-projecting.ru

Аннотация

В статье рассмотрены современные подходы к управлению семантическими активами. Предложены способы управления жизненным циклом семантических активов и их повторного использования для достижения семантической интероперабельности при взаимодействии информационных систем, моделировании процессов информационного обмена, генерации семантических веб-сервисов и переводе открытых данных в связанные открытые данные (Linked Open Data). Открытая платформа коллективной работы «Центр семантической интеграции» − проект РЭУ им. Плеханова, направленный на поддержку исследований и апробацию предлагаемых методов на базе принципов Model Driven Architecture (архитектуры, управляемой моделью, MDA).

1 Введение

Вопросам применения семантических методов интеграции данных из гетерогенных источников посвящено значительное число научных работ [4, 8, 9, 14, 21, 30, 31]. Эти методы более 15 лет успешно развиваются в Европе и США [1, 10] для обеспечения семантической интероперабельности информационных систем (ИС), например, в сфере электронного правительства.

Семантическая интеграция базируется на использовании семантических моделей данных – семантических активов (СА) [30] – глоссариев, словарей, таксономий, тезаурусов и онтологий. Построение СА выполняется, как правило, коллективами специалистов в ходе разработки ИС или моделирования той или иной предметной области. Для накопления и распространения знаний, инкапсулированных в СА, используются решения по каталогизации СА, которые обеспечивают их доступность и повторное использование. В условиях трансграничного информационного взаимодействия каталоги СА учитывают также особенности локализации. Вместе с тем существующие платформы каталогизации семантических активов не обеспечивают управление жизненным циклом (ЖЦ) СА, хотя в рамках общих рекомендаций (JOINUP [7], ADMS [23]) методология управления ЖЦ представлена.

Весьма детально управление ЖЦ метаданных регламентировано стандартами ISO/IEC 11179 [6] и соответствующим ГОСТ Р ИСО/МЭК 11179-1-2010 «Информационная технология. Регистры метаданных (РМД)» [18]. Однако они не учитывают особенностей управления семантическими активами.

Поскольку СА являются не только семантическими моделями, но и стандартами данных, целесообразно обратить внимание также на рекомендации W3C [15] по управлению ЖЦ разработки и распространения стандартов в сфере веб-технологий.

Таким образом, для решения задачи управления ЖЦ СА необходимо развитие существующих методологических разработок и действующих стандартов для учета специфики процессов разработки СА, их локализации и повторного использования.

Заметим, что потребность в повторном использовании ранее разработанных СА возрастает по мере интеграции все большего числа разнородных ИС, поскольку позволяет избежать крайне затратных и дублирующихся исследований. При переводе открытых данных в связанные открытые данные, работа экспертов направлена на применение ранее разработанных СА для придания данным смысловой окраски.

В то же время при моделировании процессов информационного обмена уже разработанных СА, как правило, не хватает и приходится решать задачу построения более детальной семантической модели обмена применительно к предметной области взаимодействующих ИС. Вместе с тем, в процессе формализации предметной области зачастую возникает «разрыв» при взаимодействии экспертов домена и разработчиков ИС, что приводит к множественным доработкам и получению неудовлетворительных результатов. Сочетание методов коллективной работы и принципов MDA − архитектуры, управляемой моделью, предложенной

Труды XVIII Международной конференции DAMDID/RCDL'2016 «Аналитика и управление данными в областях с интенсивным использованием данных», Ершово, 11-14 октября 2016

OMG (Object Management Group) [11] – призвано сократить этот разрыв.

Проект РЭУ им. Г.В. Плеханова по созданию Центра семантической интеграции (ЦСИ) включает макетирование открытой платформы коллективной работы на базе существующего международного и российского опыта, разработки и апробации методов управления семантическими активами, реализации технологий повторного использования СА для решения прикладных задач информационного взаимодействия.

2 Управление семантическими активами

В рамках европейской программы интероперабельности ISA [1] для каталогизации СА разработан стандарт ADMS (Asset Description Metadata Schema) [23], который является профилем Словаря каталогов данных DCAT (Data Catalogue Vocabulary) [29]. Причем если DCAT предназначен для упрощения взаимодействия между каталогами данных, т.е. сам каталог находится в непосредственно центре словаря, то ADMS ориентирован на СА в каталоге.

ЖЦ СА в ADMS и, соответственно, в платформе JOINUP описывается 4-мя статусам: готов, устарел, в разработке, изъят. То есть управление ЖЦ на стадии разработки СА остается за рамками платформы JOINUP.

В ЦСИ стандарт ADMS также применяется для каталогизации СА, хранения и публикации их описаний. В то же время ЖЦ СА должен учитывать стадии коллективной работы над СА, проведения экспертизы и оценки качества опубликованного семантического актива. Поэтому в рамках нашего исследования были рассмотрены как рекомендации W3C [15], так и стандарты серии ГОСТ Р ИСО/МЭК 11179 [18], включая не вошедший в российскую версию ISO/IEC 11179-6:2015 [6].

При разработке и распространении стандартов W3C используется следующие стадии ЖЦ:

Рабочий проект (WD) – документ, который W3C опубликовал для рассмотрения сообществом, в том числе членами W3C, сообществами и другими техническими организациями. Рабочие проекты не обязательно представляют собой консенсус Рабочей группы, и не подразумевают какого-либо одобрения W3C.

Кандидат на рекомендацию (CR) – документ, который удовлетворяет техническим требованиям Рабочей группы и прошел стадию экспертного обсуждения в сообществе.

Предлагаемая рекомендация (PR) – документ, который по качеству соответствует требованиям W3C и был принят директором W3C.

Рекомендация W3C (REC) – спецификация, набор руководящих принципов или требований, которые, после проведения обсуждения получили одобрение членов W3C и директора W3C.

Следует подчеркнуть, что стандарты в сфере веб-технологий рассматриваются в W3C как единый документ. В ЦСИ такой подход может быть принят только для управления описаниями семантических активов. Однако для решения задач применения и повторного использования семантических активов, которым посвящен раздел 3 данной статьи, только описаний («карточек») СА недостаточно – требуется публикация и распространение содержания семантического актива, его элементов. ISO/IEC 11179-6:2015 регулирует процесс регистрации регистров метаданных и входящих в них элементов (записей регистра) в соответствии со следующими стадиями ЖЦ:

Предпочтительный стандарт (Preferred Standard). Регистрирующий орган подтвердил, что запись в регистре предпочтительна для использования в сообществе пользователей регистра метаданных.

Стандарт (Standard). Регистрирующий орган подтвердил, что запись в регистре необходимого качества и имеет широкий интерес для использования в сообществе пользователей регистра метаданных.

Надлежащего качества (Qualified). Регистрирующий орган подтвердил, что обязательные атрибуты метаданных являются полными, обязательные атрибуты метаданных соответствуют применяемым требованиям к качеству.

Зарегистрирован (Recorded). Регистрирующий орган подтвердил, что все обязательные атрибуты метаданных заполнены.

Кандидат (Candidate). Запись в регистре предложена для прохождения регистрации.

Не завершен (Incomplete). Отправитель хочет информировать сообщество, которое использует регистр метаданных, о существовании новой записи в регистре.

Изъят (Retired). Регистрирующий орган одобрил для записи в регистре, что ее больше не рекомендуется использовать в сообществе пользователей регистра метаданных или, что она больше не должна быть использована.

Замена (Superseded). Регистрирующий орган одобрил для записи в регистре, что ее больше не рекомендуется использовать в сообществе пользователей регистра метаданных, и что теперь предпочтительно использовать заменяющую запись в регистре.

Применение этого стандарта в ЦСИ требует организации регистрирующего органа, поэтому должно быть отнесено на более поздние этапы развития Центра. В ЦСИ представлены не только описания семантических активов, собранные в каталог, но и собственно содержание СА – набор элементов и их свойств. Поэтому в ЦСИ предлагается реализовать как ЖЦ описания («карточки») СА, т.е. записи в ADMS каталоге, так и ЖЦ содержания СА. Необходимость такого деления состоит в том, что описание актива – рассматривает СА как целую, неделимую единицу, а содержание СА может разбиваться на части (ветки, разделы,

наборы, элементы), для каждой из которых может исполняться отдельный рабочий процесс.

Для описания СА может быть использован имеющий широкое распространение и признание в веб-сообщества ЖЦ W3C, поскольку он учитывает такие особенности ЦСИ, как привлечение широкого круга специалистов, разнообразие предметных областей СА, использование СА в Semantic Web и обмене информацией с использованием web-технологий. С организационной точки зрения важно отметить, что роль «Директора ЦСИ» не требуется, решение должно приниматься экспертным сообществом с использованием механизмов голосования, оценок, рейтингов и др. Семантические активы могут быть классифицированы по источникам управления. Внешние СА (например, EuroVoc) разрабатываются и развиваются за пределами ЦСИ, внутренние СА – непосредственно в ЦСИ. Внешние активы представляются в ЦСИ своими описаниями, однако в рамках ЖЦ внешних СА (например, в процессе обсуждения или экспертной оценки) могут возникнуть рекомендации по их загрузке для перевода и повторного использования. Таким образом, инициируется появление внутреннего СА, например, русской ветки EuroVoc. Внутренние СА могут создаваться участниками ЦСИ (методами ручного ввода или автоматизированной загрузки из внешних программ), а также возникают при повторном применении СА – в рамках моделирования процессов информационного обмена и при переводе открытых данных в связанные.

В ходе работы над содержанием СА (наполнение, перевод) могут возникать этапы ЖЦ, отражающие стадии коллективной работы участников рабочей группы. Например: выбор элементов СА (стадия «подготовка»), назначение заданий (стадия «перевод»), контроль исполнения заданий (стадия «проверка») и оценка качества (стадия «принят»).

Для этапов ЖЦ описания СА в соответствии со схемой [23, 28] используется свойство *«adms:status»* класса *«adms:Asset»*. А для этапов ЖЦ содержания СА − свойство *«adms:status»* класса *«adms:AssetDistribution»*

Важным этапом является экспертная оценка, уровень которой выбирается в зависимости от интереса к СА: (1) валидация в рамках рабочей группы, (2) публичная экспертиза участников сообщества, (3) независимая экспертиза, проводимая экспертами предметной области.

Переходы между этапами ЖЦ описания актива доступны только после окончания выполнения этапов ЖЦ содержания актива (рис.1). Таким образом объединяются ЖЦ описаний СА и ЖЦ процессов, выполняемых над содержимым СА. Например, невозможно будет получить состояние «Рекомендация», если ещё выполняется процесс «Экспертная оценка». При нахождении описания актива на том или ином этапе ЖЦ для изменения содержания актива могут запускаться только процессы доступные на этом этапе. Например, «Перевод» содержания актива может выполняться только, если описание актива находится на этапе «Рабочий проект».

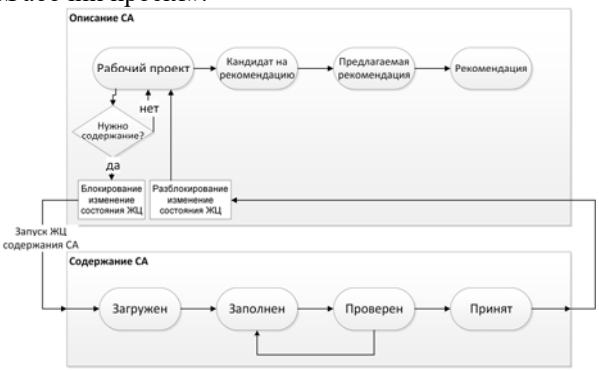


Рисунок 1 ЖЦ описания и содержания семантического актива

С учетом особенностей семантических активов и процессов коллективной работы над их содержанием, управление СА должно происходить в рамках ЖЦ описания СА, представленного в таблице 1.

Решение задач повторного использования СА и их элементов накладывает ограничение на удаление/снятие с публикации. Кроме очевидных механизмов уведомления участников, использующих СА об изменениях, должны быть поддержаны и методы «вывода из использования». Исследование показало, что существующие каталоги СА также обеспечивают именно их накопление, то есть параллельно могут существовать несколько версий СА. Это означает, что при интеграции гетерогенных систем для развития семантического ядра обмена должно быть обеспечено сохранение знаний об изменениях семантики, установление связей между «старыми» и «новыми» СА.

Например, СА в состоянии «Рекомендация» может быть снят с публикации или определен как устаревший. При этом он остается «Рекомендацией», т.е. состояние «Рекомендация» является финальным для текущей версии актива и не может быть изменено. Используя свойство [23, 28] *«owl:versionInfo»* класса *«adms:Asset»* можно установить флаг со значением «устаревший» и связь с указанием более свежей версии (свойство *«adms:last»* класса *«adms:Asset»*). Это позволяет обеспечить использование ранее разработанных СА и реализовать явные правила преобразований СА из устаревших в актуальные версии.

3 Применение семантических активов

(Повторное) использование семантических активов является базовым принципом, обеспечивающим достижение семантической интероперабельности. В проекте ЦСИ ведется исследование, апробация и развитие методов разработки и (повторного) использования СА для перевода открытых данных в связанные открытые данные, моделирования схем данных и информационного обмена при интеграции ИС, на

основе принципов коллективной работы экспертов, повторного использования СА и MDA.

Таблица 1 Этапы ЖЦ СА

Этап ЖЦ описания СА	Процессы работы над содержанием СА	Комментарий
Рабочий проект (РП)	Публикация описаний внешнего СА. Загрузка внешних СА. Создание внутреннего СА Перевод. Внесение изменений. Экспертная оценка. Удаление/снятие с публикации.	Основное наполнение содержания актива (загрузка внешнего СА, создание нового СА, перевод и т.д.) и уточнение описания СА (наполнение полей, классификация, связи с другими активами).
Кандидат на рекомендацию (КР)	Внесение изменений Экспертная оценка Удаление/снятие с публикации	Основная экспертная оценки и внесение косметических изменений
Предлагаемая рекомендация (ПР)	Экспертная оценка Удаление/снятие с публикации	Решение, что СА может быть рекомендован к использованию в ЦСИ.
Рекомендация (РЕК)	Удаление/снятие с публикации Создание новой версии	Могут приниматься решения о создании новой версии или удалении актива из каталога.

3.1 Перевод открытых данных в связанные

Концепция открытых данных (Open Data, OD) декларирует доступ к информации не просто в визуальных, но и в машиночитаемых форматах [17], что обеспечивает возможность ее многократного использования.

За годы продвижения принципа открытости сформирован значительный пул OD, однако практика показывает, что такой способ публикации не обеспечивает возможности использования данных в полной мере. Более того, анализ и визуализация открытых данных зачастую может выполняться только после дополнительной обработки или «ручным способом». Это связано с тем, что требования ни к структурированию и гармонизации данных, ни к описывающим их метаданным, в том числе связи с контекстом и терминами предметной области, в OD не предъявляются.

Большой интерес для мирового сообщества представляет концепция связанных открытых данных (Linked open data, LOD), которая использует [2] срез технологий Semantic Web для обеспечения совместимости, облегчения многократного использования и комбинирования данных. Семантическое связывание данных позволяет объединять данные из различных источников, не прибегая к созданию специализированных программ, и предоставлять доступ к этим данным.

LOD обеспечивают децентрализацию данных, помогают в работе с большими объёмами данных [2, 3]. Способы применения LOD для решения задач интеграции данных в предметных областях с интенсивным использованием данных (Data Intensive Domains) неоднократно рассматривались коллективом авторов под руководством В.А. Серебрякова, например, [21].

Исследуемые в рамках ЦСИ методы перевода открытых данных в LOD основываются на четырех сформулированных Тимом Бернесом-Ли принципах Semantic Web [2] и на дополнительном **пятом принципе** – повторного использования семантических ресурсов. В рамках проекта ЦСИ ведется исследовательская работа по апробации и развитию методов перевода открытых данных в связанные открытые данные (OD2LOD) путем формирования метаданных на основе использования СА.

В процессе перевода OD2LOD должны быть реализованы следующие этапы: (1) подготовить данные (загрузить, нормализовать, «связать»), (2) провести их верификацию и (3) получить сформированный в рамках стандарта пакет LOD, готовых для (4) публикации и (5) последующего использования в процессах информационного обмена.

Для апробации выбранных методов разработаны примеры, реализующие шаги бизнес-процесса OD2LOD. В примерах использованы открытые данные, опубликованные на официальных порталах государственных открытых данных (data.gov.ru, mos.ru).

Для подготовки данных в части загрузки и нормализации используется инструментарий Open Refine [26], а для связывания с семантическими ресурсами апробируется платформа Silk [27]. Для хранения моделей и данных применяется Apache Jena [24]. На этапе публикации особенно важным является применение средств визуализации LOD, исследование которых сегодня активно поддерживается международным научным сообществом [5].

При отсутствии готовых решений и/или при необходимости их интеграции ведется разработка собственных приложений ЦСИ для повторного использования СА для OD2LOD и выполнения всех этапов перевода.

Таким образом, в ЦСИ будут реализованы:

- Единая точка доступа – все необходимые инструменты собраны в одном месте, имеют общий интерфейс.
- Инструменты связывания – позволяют снизить стоимость самой трудоемкой операции по подготовке данных.
- Инструменты доступа к данным – позволяют получать данные из различных источников.
- Инструменты визуализации – позволяют показать данные в различных представлениях.
- Повторное использование моделей – возможность использовать или доработать имеющиеся модели под свои нужды вместо разработки новых.
- Хранение данных – возможность автоматически настраивать хранилище и хранить большие объемы данных, снижая риски утери первоначального источника.

3.2 Моделирование схем информационного обмена

Гетерогенная среда информационных систем в областях с интенсивным использованием данных, которые представляют ценность для других потребителей, диктует необходимость организации взаимодействия и моделирования схем информационного обмена.

Интеграция данных на семантическом уровне поддерживает унифицированное представление данных на основе семантических свойств в контексте единой онтологии предметной области [22]. Для достижения этого должны использоваться, в том числе повторно, семантические активы. Преимущество семантического подхода к интеграции ИС перед традиционным убедительно показано в [30].

Теоретические исследования описывают методы семантической интеграции [21, 22, 30, 31], однако на практике их применение затруднено. Несмотря на то, что «семантическое моделирование стало предметом исследований начиная с конца 1970-х годов» [4], проблема гетерогенных семантик данных, которые возникают как внутри систем, так и при обмене между ними, остается актуальной [8].

На 7-й международной конференции GRID 2016 в Дубне авторы статьи представили результаты[1] исследования существующих подходов к семантической интеграции и сформулировали ряд проблем, на решение которых направлен проект ЦСИ. В первую очередь это проблема организации взаимодействия между специалистами предметной области (домена) и ИТ-специалистами: моделирование информационной системы и процессов взаимодействия в гетерогенной среде должно опираться на знания о предметной области, ее объекты и связи между ними, то есть должно быть доступно и понятно всем участникам вне зависимости от сферы их компетенции.

Необходимость разработки специальных модулей импорта/экспорта данных возникает при интеграции ИС, поскольку схемы данных различных ИС обычно не соответствуют друг другу, причем источники данных и их получатели используют различные контексты. Неявные допущения, существующие в каждом источнике, должны быть явно описаны и использованы, чтобы исключить конфликты при сопоставлении и совместном использовании данных из этих систем [9].

Современные подходы к интеграции данных в основном ориентированы на использование XML (eXtensible Markup Language) форматов обмена данными, для описания которых используются XSD схемы (XML Schema definition). Наиболее популярные на сегодняшний день подходы к построению XSD схем используют модели на логическом уровне интеграции.

При этом построение моделей, как правило, проводится IT-специалистом, не имеющим достаточных знаний в предметной области. Моделирование происходит на конечных этапах разработки ИС в рамках платформо-зависимой реализации. Итоговый результат заказчик/потребитель может увидеть только после окончания разработки. Соответственно, такие модели семантических связей не содержат и машинное «понимание» остается невозможным.

Применение семантических технологий для реализации задач информационного обмена, в том числе семантической шины [32], требует участия в моделировании специалистов домена. Однако ориентация инструментов разработки на ИТ-специалистов определяет высокий порог вхождения экспертов домена в процессы разработки схем обмена данными (и ИТ-систем в целом).

Одной из основных задач ЦСИ является организация совместной коллективной работы экспертов домена и ИТ-специалистов. Последовательный переход от бизнес-требований к технической реализации в ЦСИ базируется на принципах архитектуры управляемой моделью (Model Driven Architecture, MDA) [11] и (повторном) использовании СА. MDA дает возможность сохранить инвестиции, сделанные в разработку бизнес-логики даже при смене технологических платформ.

Применение этого подхода в ЦСИ позволяет использовать СА при реализации схем информационного обмена и обогащении существующих данных семантическим описанием. Примером такого обогащения является установка связи каждого элемента XSD-схемы (в частности, WSDL описании web-сервиса), который является семантическим свойством данных в предметной

[1] https://indico-new.jinr.ru/indico/contributionDisplay.py?sessionId=16&contribId=62&confId=85

области, с СА или его элементами, представленными в ЦСИ. В описании XSD схемы эта информация сохраняется в элементе схемы *«xsd:documentation»*.

Основой разработки СА и их применения для решения прикладных задач является реализация 4-х уровневой иерархии метамоделей (M3-M0) [12, 13, 20]. Она обеспечивает целостность моделей, создаваемых и предоставляемых в ЦСИ, следующим образом:

- в ядре платформы ЦСИ реализуется фабрика моделей (т.н. мета-объектное средство, Meta-Object Facility, MOF) – уровень M3;
- в состав ЦСИ включается метамодель UML – уровень M2;
- для представления модели СА выбрана нотация UML (Unified Modeling Language) – уровень M1;
- СА являются экземплярами классов из моделей СА – уровень M0.

В процессе разработки СА, используя уже существующие в ЦСИ элементы СА, «предметник» должен иметь возможность сформировать структуру СА в визуальном режиме. В результате, автоматически или с минимальными уточнениями, создается общая платформо-независимая модель (Platform Independent Model, PIM) [11]. А также обеспечивается её связь с вычислительно-независимой моделью (Computation Independent Model, CIM).

Готовая PIM дополняется контекстными компонентами (например, связанными СА), формируя на выходе контекстную PIM, которая затем трансформируется в платформо-зависимую модель (Platform Specifical Model, PSM) реализуемой платформы. При необходимости замены платформы может быть проведена повторная генерация PSM на основе уже имеющейся PIM.

CIM, PIM и PSM модели могут быть визуализированы в нотации UML или разработаны во внешнем UML-редакторе, а затем загружены в ЦСИ в XMI формате. На основе PSM генерируются платформенные компоненты (создаются базы данных, схемы XSD и т.п.). Схемы XSD, обогащённые связью с СА, используются в реализации семантических сервисов (Semantic Web Services).

Такой подход позволяет в перспективе включить в ЦСИ и другие метамодели, например, CWM (Common Warehouse Metamodel). А загруженные в ЦСИ модели предметных областей, справочные модели и базовые словари, такие как ядро национальной модели обмена информацией США NIEM (National Information Exchange Model) [10], базовые европейские словари (Core Vocabularies) [25], российские справочники и классификаторы, могут быть использованы для построения схем информационного обмена и обеспечения семантической интероперабельности ИС.

4 Заключение

В статье рассмотрены подходы к управлению жизненным циклом семантических активов и возможности применения СА. Определена необходимость объединения ЖЦ описания СА и ЖЦ содержания СА, с учетом их особенностей, в том числе повторного использования. Для ЖЦ описания СА предложено применение методологии W3C, расширенной процессами работы над содержанием СА.

Приведены предложения по реализации перевода открытых данных в связанные открытые данные с использованием СА. Сформулированы основные функции процессов OD2LOD для их дальнейшей реализации в ЦСИ.

Рассмотрен подход к устранению «разрыва» при взаимодействии экспертов домена и разработчиков ИС в процессе семантического моделирования предметной области на основе сочетания методов коллективной работы, повторного использования СА и принципов MDA.

Проект РЭУ им. Г.В. Плеханова по созданию Центра семантической интеграции представлен как площадка для апробации методов и инструментария, разрабатываемых в ходе проведения научно-исследовательской работы. Создаваемый макет открытой платформы коллективной работы ЦСИ будет обеспечивать каталогизацию СА, управление их жизненным циклом, а также практическое применение СА для решения задач информационного взаимодействия.

При проведении прикладных исследований с использованием ЦСИ научные коллективы с привлечением молодых ученых и студентов получат возможность семантического моделирования различных предметных областей, создания глоссариев, таксономий, тезаурусов и онтологий для применения в исследовательских проектах, систематизации научных материалов и машинной обработке научных текстов на русском языке.

ЦСИ как семантическая база знаний, обеспечивающая возможность многократного повторного использования СА, может эффективно использоваться в областях с интенсивным использованием данных, таких как:

- Электронное правительство. Обеспечение коллективной работы разноплановых специалистов по созданию семантических моделей государственных услуг и организации межведомственного взаимодействий для обеспечения семантической интероперабельности в российском ЭП;
- Анализ данных. Новый уровень аналитики по связанным данным из разных источников, визуализация и сопоставление связанных показателей.
- Семантическая интеграция информационных систем. Владельцы ИС и коллективы

разработчиков получают возможность моделирования предметной области ИС, построения схем информационного обмена с использованием федеративных моделей данных.

Литература

[1] Annex 2 to the Communication from the Commission to the European Parliament, the Council, the European Economic and Social Committee and the Committee of Regions 'Towards interoperability for European public services' EUROPEAN COMMISSION Bruxelles, le 16.12.2010 COM(2010) 744 final http://ec.europa.eu/isa/documents/isa_annex_ii_eif_en.pdf

[2] T. Berners-Lee. Linked Data – Design Issues http://www.w3.org/DesignIssues/LinkedData.html

[3] Chris Bizer, Tom Heath, Tim Berners-Lee: Linked Data: Principles and State of the Art, Arpil 2008, http://www.w3.org/2008/Talks/WWW2008-W3CTrack-LOD.pdf

[4] C. J. Date, An Introduction to Database Systems (8th Edition). Pearson Education Inc., 2004, p. 1024, ISBN 0-321-18956-6

[5] Jiˇrí Helmich, Jakub Klimek, and Martin Neˇcaský, Visualizing RDF Data Cube using the Linked Data Visualization Model, in The Semantic Web: ESWC 2014 Satellite Events, Anissaras, Crete, Greece, Springer Verlag, ISBN: 978-3-319-11955-7, ISSN: 0302-9743, pp. 368-373, 2014

[6] International Standard ISO/IEC 11179-6:2015, Information technology — Metadata registries (MDR) — Part 6: Registration

[7] Joinup, European Commission, https://joinup.ec.europa.eu/asset/page/practice_aids/what-semantic-interoperability

[8] Madnick S., Gannon T., Zhu, H., Siegel M., Moulton A., Sabbouh M., Framework for the Analysis of the Adaptability, Extensibility, and Scalability of Semantic Information Integration and the Context Mediation Approach, Massachusetts Institute of Technology Cambridge, MA, USA, 2009

[9] Madnick, S.E., & Zhu, H., Improving data quality through effective use of data semantics, Data and Knowledge Engineering, 59(2), 2006, p.460-475

[10] National Information Exchange Model, https://www.niem.gov

[11] Object Management Group, MDA Specification, http://www.omg.org/mda/specs.htm

[12] Object Management Group, Meta Object Facility (MOF) Specifcation, 2000, http://www.omg.org/spec/MOF/2.5/

[13] Iman Poernomo, The Meta-Object Facility Typed, SAC'06 April 23-27, 2006, Dijon, France, http://calcium.dcs.kcl.ac.uk/1259/1/acm-paper.pdf

[14] Walaa S. Ismail, Mona M. Nasr, Torky I. Sultan, Ayman E. Khedr, Semantic Conflicts Reconciliation as a Viable Solution for Semantic Heterogeneity Problems, (IJACSA) International Journal of Advanced Computer Science and Applications, Vol. 4, No.4, 2013

[15] W3C Consortium Process Document - http://www.w3.org/Consortium/Process/

[16] Акаткин, Ю.М., Дрожжинов В.И., Ясиновская Е.Д., Что такое система межведомственного взаимодействия NIEM, 2014, http://www.cnews.ru/reviews/new/ikt_v_gossektore_2014/articles/chto_takoe_sistema_mezhvedomstvennogo_vzaimodejstviya_niem/

[17] Бегтин И., Дубова Н., О данных в открытую, Открытые системы № 03, 2015, http://www.osp.ru/news/articles/2015/09/13045074/

[18] ГОСТ Р ИСО/МЭК 11179-1-2010, Информационная технология. Регистры метаданных (РМД)

[19] Кузнецов М.Б., Трансформация UML-моделей и ее применение в технологии MDA, 2016, http://citforum.ru/SE/project/uml_mda/

[20] Марков Е., Архитектура, управляемая моделью, 2005, http://citforum.ru/gazeta/13/

[21] Малахов Д.А., Серебряков В.А., Теймуразов К.Б., Шорин О.Н., Интеграция библиографических данных в Linked Open Data, Труды 16-й Всероссийской научной конференции «Электронные библиотеки: перспективные методы и технологии, электронные коллекции» — RCDL-2014, Дубна, Россия, 13–16 октября 2014 г.

[22] Серебряков В.А, Семантическая интеграция данных, 2012, http://sp.cmc.msu.ru/proseminar/2012/serebryakov.2012.04.20.pdf

[23] Спецификация и описание Asset Description Metadata Schema, 2016 https://joinup.ec.europa.eu/asset/adms/home

[24] Спецификация и описание Apache Jena, http://jena.apache.org/

[25] Спецификация и описание EU Core Vocabularies, 2015, https://joinup.ec.europa.eu/node/145983

[26] Спецификация и описание Open Refine, http://openrefine.org/

[27] Спецификация и описание Silk Framework, http://silkframework.org/

[28] Спецификация и описание W3C Asset Description Metadata Schema (ADMS), август 2013, http://www.w3.org/TR/vocab-adms/

[29] Спецификация и описание W3C Data Catalogue Vocabulary,2014, http://www.w3.org/TR/2014/REC-vocab-dcat-20140116/

[30] Черняк Л., Интеграция данных: Семантика и синтаксис, Открытые системы №10, 2009, http://www.osp.ru/os/2009/10/11170978/

[31] Шибанов С. В., Яровая М. В., Шашков Б. Д., Кочегаров И. И., Трусов В. А., Гришко А. К. Обзор современных методов интеграции данных в информационных системах // НиКа. 2010. №. URL: http://cyberleninka.ru/article/n/obzor-sovremennyh-metodov-integratsii-dannyh-v-informatsionnyh-sistemah

[32] Ульянов Д., Проект семантической шины semap, апрель 2007, https://dulanov.wordpress.com/2007/04/17/proekt-semanticheskogo-menadgera-semap/ T. Berners-Lee. Linked Data – Design Issues http://www.w3.org/DesignIssues/LinkedData.html

Management and (re)use of semantic assets for information sharing

Yury Akatkin, Elena Yasinovskaya, Mikhail Bich, Andrey Shilin

The article represents current research results gained at The Center of Semantic Integration (Plekhanov Russian University of Economics). It shows the approach to the management of semantic assets and their (re) use for information sharing, data modelling based on Model Driven Architecture, semantic web services generation and transformation of open data to Linked open data.

Sharing research facilities data in common data infrastructures

© Vasily Bunakov © Alistair Mills

Science and Technology Facilities Council,
Harwell, United Kingdom
vasily.bunakov@stfc.ac.uk , alistair.mills@btinternet.com

© Piotr Oramus

AGH University of Science and Technology,
Kraków, Poland
oramus@student.agh.edu.pl

Abstract

The work describes the collaboration between a large experimental research facility and emerging national and cross-national data infrastructures, with the purpose of sharing experimental data and making it findable in common multi-disciplinary data catalogues.

1 Introduction

Many of the major centres of scientific research provide both the instruments for the research, and the infrastructure for storing and processing data. This is typical for large research facilities like synchrotrons, neutron sources, powerful lasers that grant timeslots to visitor scientists for their specific investigations and provide infrastructure for data collection and preservation. Generally, scientists work on the science and facility IT engineers work with the data; this leads to a requirement that these two groups collaborate. Another requirement for collaboration comes from the emerging e-infrastructures that transcend institutional and national borders and research disciplines.

Although research facilities make the data available, they do not provide a large range of access methods. The purpose of our work was to provide an industry standard protocol for accessing the data so that a large number of researchers can find the records about datasets produced by research facilities and access them easily.

New routes to existing data and metadata are important as in the last decade the number of data sources in Europe has increased enormously. It is no longer viable for most researchers to track all of the data which are relevant to their investigations, so data discovery services provided by a cross-discipline infrastructure are essential. Our work is an example of a productive collaboration between a discipline-specific data centre – ISIS neutron and muon facility [3] that is a part of a wider landscape of similar neutron and photon facilities in Europe [9] – and EUDAT e-infrastructure [1] using popular metadata standards and protocols.

Proceedings of the XVIII International Conference «Data Analytics and Management in Data Intensive Domains» (DAMDID/RCDL'2016), Ershovo, Russia, October 11 - 14, 2016

2 Use case description

EUDAT has developed several services, namely:

- B2SHARE – a data publishing service;
- B2SAFE – a secure and reliable replication service;
- B2FIND – a data discovery service (data catalogue);
- B2STAGE – a data delivery service for the rapid delivery of large volumes of data towards high-performance computing;
- B2ACCESS – user authentication service used by some of the above services.

EUDAT services are deployed centrally by project participation organizations with free registration and access for researchers, or the services can be deployed by interested parties in their own environment as all the software in support of these services is open source. We have focused on using the centrally deployed instance of EUDAT B2FIND [8] which consumes records delivered by data providers using OAI-PMH [2], maps them to its own metadata schema, and publishes them in a common data catalogue. The OAI-PMH specification is straightforward and allows the use of different metadata schemas; however, within a single metadata schema, quite different interpretations of metadata elements are possible; EUDAT always negotiates the meanings of metadata elements with the data provider.

The data provider in our case is the ISIS neutron and muon source [3] that collects data during scientific investigations, and that catalogues the data using the ICAT software platform [4]. ISIS has a data management policy [7] that provides public access to most of its publicly funded data at the end of an embargo period of three years. The ISIS policy requires that users of the data register with ISIS, and ISIS records their activity. Registration is free, but the management of ISIS wants to be aware of the use of its data when assessing the impact of the facility.

The work of providing ISIS data in EUDAT involved the following steps:

- evaluation of the available technology;

- building the metadata harvester;
- mapping the domain-specific metadata to a more popular schema;
- mapping the data provided by the service end point to the requirements of B2FIND;
- provision of a service end point for publishing metadata;
- liaison with EUDAT B2FIND for testing the end point and harvesting the data records.

There were two main challenges to address during implementation. The first challenge was the mapping of the metadata: from ISIS to OAI, then from OAI to B2FIND. The second challenge was to avoid compromising the data policy set by ISIS.

The first challenge was technical and required careful programming as well as discussions with specialists knowledgeable of the metadata models for both the data provider and the data consumer.

The second challenge required access to the data records so that the harvester could collect them. In order to get this access, ISIS provides suitable credentials, and it was decided to restrict harvesting to the data records with persistent identifiers in DataCite [10], as this implies that the records are not withheld by ISIS under its data embargo policy.

3 Technology stack and metadata mapping

We chose the Qualified Dublin Core (QDC) metadata schema [6] to represent the data from ISIS. This schema is well known, has a large user base and is one of the schemas recognized by the EUDAT B2FIND metadata mapping interface. The data from ISIS is well structured but it is in a schema that is not supported by the EUDAT B2FIND. The main purpose of B2FIND is data discovery rather than the harmonization of metadata schemas. Table 1 presents the mapping from ICAT metadata schema to QDC and to EUDAT B2FIND schema. This mapping is essential for the semantics of the ISIS data records once they are harvested by EUDAT.

We then developed software that harvests the data records from the ISIS data catalogue, maps them to the QDC schema and passes them to the OAI-PMH server that implements a popular standard for automatic data harvesting [2] required by EUDAT B2FIND ingest mechanism. We considered several implementations of OAI-PMH, and chose a Java implementation called jOAI [5] as it is mature, well documented and widely used. The data records acquisition component is a Python wrapper to ISIS ICAT API.

The resultant technology stack is presented by Figure 1. The bottom layer is a domain-specific data catalogue supported by the research facility (ISIS); the top layer is a multidisciplinary data catalogue supported by a common data infrastructure (EUDAT); the middle layers are components that enable a transformation from a domain-specific implementation to a common data discovery service.

We have stored the software which was developed in this project in a public repository, so that others can examine it for the details [12]. The software is modest in size, and can be easily deployed on a small computer. The computer has to execute a script once per hour to find new data, and it has to run a jOAI server continuously.

Table 1 Mapping from ICAT metadata to Dublin Core and EUDAT B2FIND

ICAT field	QDC term	B2FIND field
Investigation ->doi	dc:identifier	-
Investigation ->title	dc:title	title
Investigation ->summary	dc:description	notes
Instrument ->fullName Investigation ->name InvestigationP arameter->name (multiple)	dc:relation	tags
“dx.doi.org/” + Investigation- >doi	dcterms:referen ces	URL
User->fullName	dc:creator	author
-	-	spatial
Name of the organization (as a literal)	dc:contributor	maintaine r
Description of a facility (as a literal)	dc:subject	disciplin e
-	-	Publicati onYear
Investigation- >releaseDate	dcterms:issued	Publicati onTimesta mp
en	dc:language	Language
Facility->name Facility ->fullName Facility->url	dc:publisher	Origin
DatafileFormat ->name DatafileFormat ->type DatafileFormat ->version DatafileFormat ->description	dc:format	Format
Facility title (as a literal)	dc:relation	Geographi cDescript ion
Web link (URL) to ISIS Data Management Policy	dc:rights	Rights
-	dc:relation	Project
Country code (as a literal)	dc:relation xsi:type= ”dcterms:ISO316 6”	Country
-	-	Geographi cCoverage
Investigation ->startDate	dcterms:tempora l	TemporalC overage: BeginDate
Investigation ->endDate		TemporalC overage: EndDate

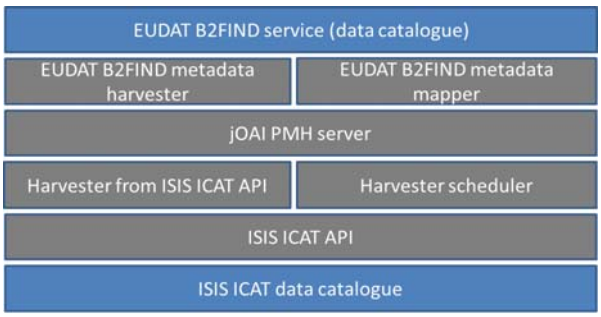


Figure 1 Technology stack for the facility-specific data discovery service

For the published information to be visible, it is necessary to register the jOAI server with a discovery service, such as B2FIND. The operation of the discovery service is the responsibility of a third party such as EUDAT.

The essential flow of work of the software is the following:

- Once per hour, the software connects to the ICAT and requests details of any new records to publish. A suitable record has a Digital Object Identifier and a Release Date since the last time the software was run;
- For each record identified, the software serializes the record as a QDC object and passes it to the jOAI publisher;
- Once per hour, the jOAI publisher checks for new objects and publishes them.

In this way, new records created by the data owner, are generally available within two hours, with no manual processing. No changes, other than configuration, are required to the ICAT server, the jOAI server or the discovery service. For the owner of the data, the additional processing required to provide this service is negligible. For the owner of the discovery service, the additional processing is negligible.

4 Data discovery use case

The services that we have developed in course of this work support the following data discovery use case. In order to find data, the researcher uses a Google-style free string search in the B2FIND data catalogue [8], and locates candidate datasets of interest. This is similar to using any search engine, except that B2FIND is likely to be more relevant as it has a harvesting policy which ensures that it searches a known set of sources; many of the sources known to B2FIND are of little general interest, and are not harvested by general purpose search engines.

Having received search results, the user selects one of the candidates located by B2FIND. B2FIND presents more information about the chosen candidate. In the case of an ISIS record, this information includes the DOI assigned to the dataset by the DataCite service [10]. The DOI link references a web landing page supplied by the ISIS facility; the landing page contains an actionable link that allows the user to get the data collected during the experiment, with the user access to the actual experimental data regulated by a facility data management policy – which in the case of ISIS is a liberal policy which encourages research data reuse [7].

Apart from its usage in EUDAT B2FIND, the OAI-PMH endpoint for ISIS ICAT and the appropriate metadata mapping are being tested for the new Research Data Discovery Service (RDDS) which is a national UK initiative similar to EUDAT B2FIND but with a different scope of research data records collected [11]. RDDS is going to become another public channel for the dissemination of experimental data collected by the ISIS facility, along with EUDAT, DataCite and research papers that cite data DOIs. Figure 2 represents the flow of data records and data persistent identifiers between different services of a common data discovery ecosystem.

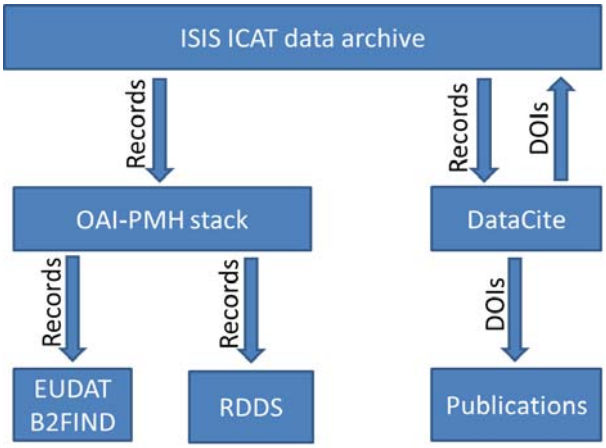


Figure 2 Data records and data DOIs flow

After a period of testing with a few harvesting e-infrastructures, the OAI-PMH stack has the potential to become part of the ICAT software distribution [4] that is used by other neutron and photon facilities in Europe. This should make it easier for other facilities to supply their data records to data discovery portals. It was not possible during the course of the project described in this paper to assess the impact of this work on the various stakeholders. However, the existence of projects such as EUDAT and RDDS and their active collaboration with this project supports our belief in the need for such projects. As we continue to work in this area, we will learn more about the needs of the stakeholders, and change our implementation to support those needs.

Conclusion

We considered the effort to implement the OAI-PMH endpoint and supply data records in e-infrastructures worthwhile for the following reasons:

- large research facilities such as ISIS have an interest in sharing data; it may be a legal or policy requirement that they publish this data, especially data that is collected in a publicly funded investigation; many investigators consider that the provision of data enhances the value of their research and consider that data citation is as valuable as publication citation,

hence more routes to citable data are beneficial for researchers;

- sharing data in multi-disciplinary catalogues like B2FIND and RDDS attracts new collaborators, facilitates data reuse within a discipline, and encourages cross-discipline research;
- we are working within a community of European facilities which are adopting common standards for software and infrastructure [9]; the software developed in the course of this work and shared in GitHub [12] provides added value in the technology stack already adopted by similar research centres, which makes our solution organizationally scalable;
- other e-infrastructures can use the ISIS ICAT OAI-PMH endpoint that is now running as beta-service [13], to harvest data records for ISIS investigations with actionable links to publicly available data; metadata cross-walks need to be defined between the OAI-PMH metadata and the e-infrastructure metadata; this is similar to EUDAT, and aims to avoid semantic misinterpretation of metadata elements.

This work provides foundation IT-components and from an organizational point of view, may serve as a model for sharing data collected by large research facilities in common cross-disciplinary data infrastructures. The work is a contribution to the emerging European research data ecosystem comprising traditional research centres, common national and transnational e-infrastructures, research teams located in smaller labs in universities and industry, as well as individual researchers willing to share data. The work aims to increase the efficacy and efficiency of using the public funds allocated for research and development, by providing new routes for data publishing and data reuse.

Acknowledgements

This work is supported in part by Horizon 2020 EUDAT and the UK JISC RDDS projects, although the views expressed are the views of the authors and not necessarily of the projects.

References

[1] EUDAT: the collaborative Pan-European data infrastructure. http://www.eudat.eu

[2] Open Archives Initiative Protocol for Metadata Harvesting. https://www.openarchives.org/pmh

[3] ISIS neutron and muon research facility. http://www.isis.stfc.ac.uk

[4] ICAT project. http://icatproject.org

[5] jOAI. http://www.dlese.org/oai

[6] DCMI Metadata Terms. http://dublincore.org/documents/dcmi-terms

[7] ISIS data policy. http://www.isis.stfc.ac.uk/user-office/data-policy11204.html

[8] EUDAT B2FIND service. http://b2find.eudat.eu

[9] PaNdata initiative. http://pan-data.eu

[10] DataCite service. http://www.datacite.org

[11] UK Research Data Discovery Service. https://www.jisc.ac.uk/rd/projects/uk-research-data-discovery

[12] PMH component in ICAT GitHub repository https://github.com/icatproject-contrib/pmh

[13] ISIS ICAT OAI-PHM endpoint (beta-service). http://oai.eudat.stfc.ac.uk/oai/provider?verb=Identify

Формальная семантика языка разрешения сущностей и слияния данных и ее применение для верификации потоков работ интеграции данных

© С. А. Ступников

Институт проблем информатики Федерального исследовательского центра «Информатика и управление» Российской академии наук
Москва
sstupnikov@ipiran.ru

Аннотация

В течение всего периода развития методов и средств интеграции возникали вопросы их верификации, т.е. формальной проверки на соответствие заданным требованиям. Можно выделить следующие уровни интеграции: интеграция моделей данных, сопоставление и интеграция схем данных и собственно интеграция данных. В работе рассматривается подход определению формальной семантики высокоуровневых программ интеграции данных в языке спецификаций, поддержанном средствами формального доказательства. Рассматриваемая семантика применяется для верификации потоков работ интеграции структурированных данных. Свойства программ, подлежащие проверке, представляются в виде выражений выбранного языка спецификаций. Затем, с использованием формальных средств доказательства, спецификация, выражающая семантику конкретного потока работ интеграции данных, проверяется на соответствие необходимым свойствам. Практической целью работы является определение оснований для формальной верификации свойств потоков работ при решении задач в различных средах интеграции данных.
Работа выполнена при поддержке РФФИ (гранты 14-07-00548, 15-29-06045, 16-07-01028).

1 Введение

Разработка методов и средств интеграции данных становится в настоящее время все более актуальной в связи со значительным ростом объемов и разнообразия данных. Различные по семантике и структурированности данные могут быть необходимы при решении одной задачи в науке или промышленности. Увеличение объемов данных требует применения масштабируемых платформ распределенной параллельной обработки данных, таких, как Apache Hadoop [4]. При этом все составляющие процесса интеграции данных должны быть реализованы в виде программ над соответствующей платформой.

Можно выделить следующие уровни интеграции (от более высокого к более низкому): интеграция моделей данных [18, 19], сопоставление и интеграция схем данных [29] и собственно интеграция данных. Завершение интеграции на более высоком уровне обычно является предусловием для интеграции на более низком уровне. Необходимо отметить, что в данной работе рассматривается интеграция *структурированных* данных, т.е. данных, конформных некоторой схеме, определяющей типы (структуры) данных.

На уровне моделей данных происходит сопоставление элементов моделей (языков), на уровне схем – сопоставление типов (структур) данных и их атрибутов, на уровне собственно данных определяется, каким именно образом следует преобразовывать исходные коллекции данных, элементы коллекций и их атрибуты. Идеальным описанием процесса интеграции данных можно считать совокупность декларативных правил (подобная программе на языке Datalog), такой подход называется *обмен данными* (data exchange) [11]. Однако, на практике, процесс интеграции данных представляет собой набор операций трансформации данных, представленных в SQL или подобном ему языке, причем важным является порядок исполнения операций. Обычно интеграция данных организуется в виде потоков работ. Как в случае обмена данными, так и в случае потоков работ, интеграция данных согласована с предыдущим этапом интеграции – сопоставлением схем.

Процессы интеграции данных являются достаточно важными и сложными, и в течение всего периода развития методов и средств интеграции возникали вопросы определения их формальной семантики и *верификации*, т.е. формальной проверки на соответствие заданным требованиям. В ИПИ ФИЦ ИУ РАН данным вопросам в последние годы уделялось существенное внимание. Разрабатывались методы и средства унификации моделей данных с доказательным сохранением семантики [17, 18, 19,

Труды XVIII Международной конференции DAMDID/RCDL'2016 «Аналитика и управление данными в областях с интенсивным использованием данных», Ершово, 11-14 октября 2016

20, 35, 34], методы и средства верификации композиционного проектирования информационных систем [33, 31] (относящиеся к уровню схем). Разрабатываются продвинутые среды интеграции данных и решения задач над неоднородными информационными ресурсами, базирующиеся на Hadoop [21, 22, 36]. Краткий обзор упомянутых и родственных им работ приведен в разделе 2. Настоящая работа нацелена на развитие методов определения формальной семантики и верификации на третьем уровне – уровне данных. При этом предполагается, что необходимое определение семантики и верификация на уровне моделей данных и схем уже проведена. Предполагается также, что процесс интеграции данных представляется в виде набора операций трансформации данных на SQL-подобном языке, а не в виде декларативной Datalog-подобной программы. При этом свойства процесса интеграции в целом не выражаются явно, определяются лишь правила преобразования данных в конкретных операциях.

В процессе интеграции собственно данных обычно выделяют различные этапы: трансформация данных, разрешение сущностей (entity resolution) [23, 13] и слияние данных (data fusion) [6, 7, 10]. *Трансформация данных* подразумевает их преобразование из исходной схемы (схемы коллекции - источника данных) в целевую (единую интегрированную схему). Под *разрешением сущностей* обычно понимают выделение и связывание информации об одной и той же сущности реального мира из разных коллекций данных [23]. Под *слиянием сущностей* понимают комбинацию различных представлений одной и той же сущности реального мира в единое представление [28]. Этапы интеграции данных могут состоять из большого количества операций, и не обязательно идут в указанном порядке. Поэтому более правильно говорить, что процесс интеграции данных представляет собой поток работ, деятельности которого представляют собой отдельные операции трансформации структурированных (типизированных) данных, разрешения или слияния сущностей. Ряд обобщенных операций слияния данных выделен в работе [7].

Наряду с платформами распределенной параллельной обработки данных разрабатываются высокоуровневые языки программирования, которые могут быть использованы (или прямо предназначены) для трансформации данных, разрешения и слияния сущностей в рамках этих платформ. К таким языкам относятся, в частности, Jaql [5], Pig Latin [29], Highlevel Integration Language (HIL) [15]. Распределенное исполнение программ, написанных на таких языках, в среде Hadoop достигается путем компиляции высокоуровневых программ в программы на императивных языках (например, Java), которые, в свою очередь, исполняются с использованием средств поддержки в Hadoop вычислительной модели MapReduce [27].

Идея подхода, предлагаемого в данной статье, состоит в том, чтобы сообщить высокоуровневым программам интеграции данных семантику в некотором языке спецификаций, поддержанном средствами формального автоматического и/или интерактивного доказательства: для языка интеграции данных строится его отображение в язык спецификаций. Свойства программ, подлежащие проверке, представляются в виде выражений выбранного языка спецификаций. Затем, с использованием формальных средств доказательства, спецификация, выражающая семантику конкретного потока работ интеграции данных, проверяется на соответствие необходимым свойствам.

Для иллюстрации подхода в качестве языка интеграции данных выбран HIL – язык, разработанный компанией IBM, поставляемый в составе Hadoop-решения BigInsights [14], и используемый, например, в проектах интеграции финансовых и социальных данных [15]. HIL – это язык высокого уровня для описания сложных потоков обработки данных, включающих операции трансформации данных, разрешения сущностей и слияния данных. Данные при этом могут поступать из больших коллекций данных, представленных в различных схемах. Язык был применен, в частности, в проекте интеграции данных из социальных сетей [15], была продемонстрирована масштабируемость применения языка как по разнообразию схем данных, так и по объему данных. По сравнению с традиционными средствами построения процессов извлечения-трансформации-загрузки (ETL) данных, основанных на реляционной модели и языке SQL, язык HIL предлагает значительно более гибкие декларативные средства интеграции данных, нацеленные на разрешение сущностей и слияние данных.

В качестве языка спецификаций выбран язык Нотация абстрактных машин (AMN) [1], поддержанный средствами формального доказательства [2]. Выбор AMN мотивирован возможностями этого языка по спецификации операций трансформации данных и опытом автора по использованию AMN для определения формальной семантики. Для формализации семантики языка HIL недостаточно лишь средств верификации процессов (таких, как конечные автоматы или сети Петри) – необходимы средства верификации сложных операций (деятельностей, составляющих процессы). Вместо AMN возможно использование и других языков, предоставляющих аналогичные возможности, например, RAISE или Z.

Структура статьи выглядит следующим образом. В разделе 2 рассматриваются родственные работы по определению формальной семантики и верификации процесса интеграции данных на разных уровнях. В разделе 3 излагаются и иллюстрируются на примерах основные принципы семантического отображения языка HIL в язык AMN. В разделе 4 иллюстрируется на примере верификация свойств операций

разрешения сущностей и слияния данных. В заключении подведены итоги статьи и обозначены направления дальнейшей работы.

2 Родственные работы

Работы по определению семантики языков концептуального моделирования и моделей данных известны с конца 1990-ых гг. В работе [16] отображение типа связи модели ODMG'93 в AMN было использовано для верификации его отображения в каноническую модель данных. При отображении моделей данных должны сохраняться информация и семантика операций языка манипулирования данными.

Работы Lano, Bicarregui (например, [25]) посвящены формальному определению языка UML в языке RAL (Real-time Action Logic). Показано, как полученная семантика может использоваться для верификации преобразования UML-диаграмм (усиления инвариантов, рационализации ассоциаций, элиминации ассоциации много-ко-многим, транзитивности агрегации, преобразования интерфейса).

В работе [33] определена семантика объектной модели данных в языке AMN для верификации композиционного проектирования информационных систем: доказательства того, что композиция готовых программных компонентов может быть использована в качестве реализации абстрактной спецификации системы, подлежащей созданию.

В дальнейшем, язык AMN использовался для верификации отображения различных классов моделей данных: процессных [17], онтологических [20], массив-базированных [34], графовых [35].

Были разработаны обобщенные методы и средства унификации моделей (приведения их к единой канонической модели данных) [18, 19]. Верификация отображений моделей при этом также основывалась на представлении семантики моделей в языке AMN.

Много работ известно в области верификации *трансформаций моделей* [9]. Эти работы проводятся в рамках Движимой моделями инженерии (MDE). Под моделями в MDE понимаются как модели данных (языки), так и концептуальные схемы. Трансформация моделей – это реализация их отображения, набор правил, в совокупности определяющих, каким образом сущности исходной модели должны быть преобразованы в сущности целевой модели. Верификации подлежат такие свойства трансформаций, как *завершаемость* (процесс трансформации успешно завершается для любой правильно определенной исходной модели), *определенность* (уникальность целевой модели для заданной исходной модели и трансформации), *синтаксическая корректность* трансформации по отношению к языку трансформаций, *сохранение семантики исполнения* (трансформация исполняется в соответствии с ее спецификацией). Для формального доказательства свойства трансформаций представляются в формальных языках, различных вариантах логики первого порядка (например, Calculus of Inductive Constructions). В зависимости от мощности используемого языка, для доказательства свойств могут быть использованы автоматические средства (SAT-решатели, средства проверки моделей) или автоматизированный логический вывод.

Известны также различные подходы к определению формальной семантики трансформации данных, например, в работе [24] рассматривается формальная семантика диаграмм потоков данных в языке VDM (Vienna Development Method). В работе [32] для представления семантики процессов извлечения-трансформации-загрузки (ETL) данных используется язык LDL (Logic-Based Data Language). Целью этих работ является обеспечение перехода от концептуальной модели процессов трансформации данных к логической. В системе Clio [12] реализован подход обмена данными, когда трансформация данных представлена в виде Datalog-подобной программы с логической семантикой.

Особенность данной работы состоит в том, что формальная семантика в языке AMN сообщается высокоуровневому языку разрешения сущностей и интеграции (HIL), семантика применяется для верификации потоков работ интеграции данных, выраженных на HIL. Основные черты подхода выделены в следующих разделах.

3 Формальная семантика языка разрешения сущностей и интеграции HIL

Подход к определению формальной семантики языка HIL демонстрируется в данном разделе на примере интеграции финансовых данных, публикуемых Комиссией по ценным бумагам и биржам (SEC) США. Пример рассматривался в работах, посвященных системе интеграции Midas [8] и языку HIL [15]. Целью рассматриваемого потока работ интеграции данных (являющегося частью общей задачи интеграции финансовых данных) является формирование коллекции сущностей *Person,* представляющей информацию о лицах, управляющих ведущими компаниями США (рис. 1). Исходными для потока являются две коллекции: *InsiderReportPerson* (для краткости, *IRP*) и *JobChange*. В первой коллекции содержатся данные, извлеченные из внутренних отчетов компаний о заработной плате сотрудников. Во второй коллекции содержатся данные, извлеченные из отчетов о принятии на работу или смене позиций сотрудников компаний.

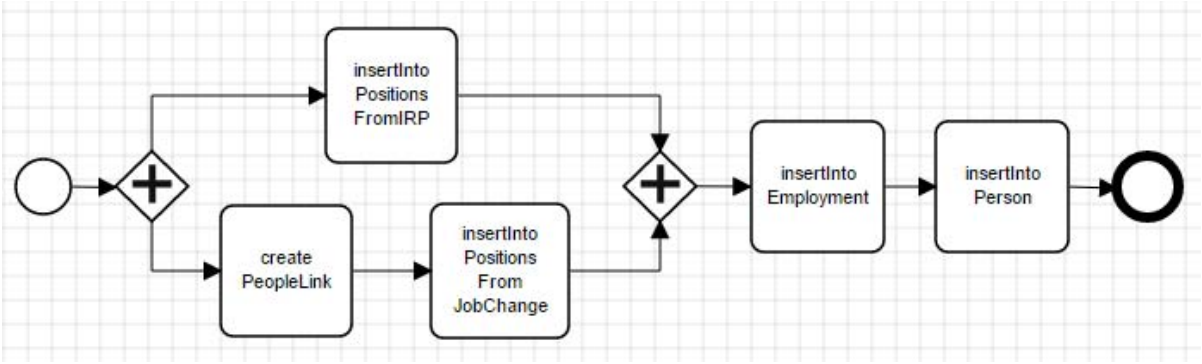


Рисунок 1 Поток работ интеграции данных об управляющих компаниями лицах

Само извлечение информации из неструктурированных данных не является предметом настоящей работы. Поток включает пять операций:

- *insertIntoPositionsFromIRP* (извлечение данных о позициях управляющих лиц из коллекции *IRP* и включение их в промежуточную коллекцию *Positions*);
- *createPeopleLink* (создание коллекции *PeopleLink* связей между управляющими лицами, упомянутыми во внутренних отчетах, и записями о смене позиций);
- *insertIntoPositionsFromJobChange* (извлечение данных о позициях управляющих лиц из коллекции *JobChange* и включение их в промежуточную коллекцию *Positions*);
- *insertIntoEmployment* (включение данных о найме сотрудников компаниями в промежуточную коллекцию *Employment*);
- *insertIntoPerson* (включение данных об управляющих лицах и их местах работы в коллекцию *Person*).

Среди рассматриваемых операций *createPeopleLink* является характерным примером разрешения сущностей, а операции *insertIntoEmployment* и *insertIntoPositionsFromJobChange* – слияния сущностей.

Следует отметить, что в языке HIL отдельным операциям трансформации, разрешения или слияния сущностей не присваиваются имена, операции выражаются в виде правил в SQL-подобном синтаксисе. Приведенные выше имена присвоены соответствующим правилам для удобства рассуждений.

Также нужно отметить, что порядок исполнения операций в языке HIL напрямую не фиксируется [15]. Этот порядок определяется при компиляции с тем ограничением, что все промежуточные коллекции, используемые операциями, должны быть материализованы, т.е. исполнены операции, осуществляющие включение данных в эти коллекции. Поток работ, изображенный на рис. 1 в графическом представлении языка BPMN, отражает неявную семантику ограничений на последовательность исполнения операций в рассматриваемом примере. Ромб со знаком «+» означает разделение потока на параллельные ветви и слияние параллельных ветвей. Например, операция *insertIntoEmployment* должна быть исполнена после операций *insertIntoPositionsFromIRP, createPeopleLink, insertIntoPositionsFromJobChange,* поскольку она использует коллекции *PeopleLink* и *Positions,* формируемые соответствующими операциями.

Семантика основных конструкций языка HIL будет определена ниже в языке AMN, основанном на теории множеств и типизированном логике предикатов первого порядка. Спецификации AMN называются *абстрактными машинами* и сочетают в себе пространства состояний и поведения машины, определенного операциями на состояниях. Ниже перечисляются и иллюстрируются на примерах основные принципы определения семантики HIL.

1. Семантика программ на языке HIL. Каждая программа на HIL представляется в AMN отдельной конструкцией вида REFINEMENT [BAMN] (такого рода конструкции допускают наиболее широкие возможности определения спецификаций). В частности, упомянутый выше пример программы представляется конструкцией *PersonFusion*:

```
REFINEMENT PersonFusion
```

2. Семантика операторов объявления типов сущностей и типов индексов. Типы сущностей и типы индексов представляются в AMN переменными абстрактных машин в разделе VARIABLES, и типизируются в разделе INVARIANT. Например, *тип сущностей* IRP, определяемый в HIL следующим образом (атрибут *name* отвечает имени персоны, *cik* – уникальному идентификатору персоны в документации SEC, *bdate* – дате рождения, *company* – имени компании, *companyCIK* – идентификатору компании, *title* – занимаемой должности, флаги *isOfficer* и *isDirector* – является ли персона служащим или управляющим лицом компании):

```
declare IRP: set [ name: string, cik: int, bdate: string;
  company: string, companyCIK: int, title: string,
  isOfficer: Boolean, isDirector: Boolean ];
```

представляется в конструкции *PersonFusion* следующим образом:

```
VARIABLES IRP, …
INVARIANT
 IRP: POW(struct(
   name: STRING_TYPE, cik: INT,
   bdate: STRING_TYPE, company: STRING_TYPE,
   title: STRING_TYPE, isOfficer: BOOL,
   isDirector: BOOL
   )) & …
```

Для типа *IRP* в разделе переменных объявляется одноименная переменная, которая типизируется в инварианте как подмножество (POW) множества всех структур (struct) – кортежей. Кортежи состоят из атрибутов, соответствующих атрибутам типа *IRP*. Между встроенными типами HIL и AMN установлено взаимооднозначное соответствие, используемое при типизации атрибутов структур в AMN. Например, тип *string* языка HIL соответствует типу STRING_TYPE в AMN, тип Boolean – типу BOOL и т.д.

Тип индекса (представляющего собой отображение из множества экземпляров типа ключа в экземпляры типа значения) *Positions*, определяемый в HIL следующим образом (ключ задается парой ⟨идентификатор персоны *cik,* имя компании *company*⟩, а значение – множеством должностей *set[title]*):

```
declare Positions:
fmap [cik: int, company: string] to set [title: string];
```

также представляется в конструкции *PersonFusion* одноименной переменной, типизируемой в инварианте:

```
VARIABLES … Positions, …
INVARIANT
Positions: struct(cik: INT, company: STRING_TYPE) +->
  POW(struct(title: STRING_TYPE))
```

Для представления отображения используется конструктор частичной функции AMN, обозначаемый символами «+->». Типы ключа и значения, как и в случае с типом сущностей *IRP*, также представляются при помощи конструкторов структур и множества подмножеств.

3. Семантика операций трансформации, разрешения и слияния сущностей. Каждому правилу языка HIL ставится в соответствие операция абстрактной машины AMN. Например, правилу создания коллекции *Empoloyment* в HIL:

```
insert into Employment …
select … from … where … ;
```

соответствует операция *insertIntoEmployment* конструкции *PeopleFusion*:

```
OPERATIONS
  insertIntoEmployment =
  SELECT … THEN … END;
```

3.1. Семантика включения данных в коллекции с типами сущностей. Примером операции такого рода является создание коллекции *PeopleLink* в HIL*:*

```
create PeopleLink as
select [cik: p.cik, docID: j.docID, span: j.span ]
from IRP p, JobChange j
match using
        rule1: normName(p.name) = normName(j.name)
check if not(null(j.bdate)) and not(null(p.bdate))
then j.bdate = p.bdate;
```

Сущности формируемой коллекции определяются в секции *select* и включают атрибуты *cik* (идентификатор персоны)*, docID* (номер документа, где встречается упоминание о смене должности персоны), *span* (место в документе, где встречается упоминание о смене должности). Данные извлекаются (секция *from*) из коллекций *IRP* и *JobChange* (происходит соединение коллекций). В секции *match* происходит предварительный отсев кортежей, полученных в результате соединения коллекций: имена персон *name* должны в обеих коллекциях совпадать с точностью до нормализации (функция *normName*). В секции *check* происходит дополнительная проверка отобранных в *match* кортежей: в случае, если в обеих коллекциях имеются данные о дате рождения персоны, даты должны совпадать. В AMN формирование коллекции *PeopleLink* осуществляется в операции *createPeopleLink:*

```
createPeopleLink =
SELECT …
THEN
PeopleLink :=
{ rr | #(pp, jj).( pp: IRP & jj: JobChange &
        rr = rec(cik: pp'cik, docID: jj'docID, span: jj'span) &
        normName(pp'name) = normName(jj'name) &
        (jj'bdate /= null_string & pp'bdate /= null_string =>
              jj'bdate = pp'bdate)  )
};
…
END
```

Переменной *PeopleLink* присваивается множество, образованное при помощи конструкции выделения множества *{rr | F(rr)}* (здесь *rr* – имя переменной, отвечающей элементу множества, *F(rr)* – формула, которая должны обращаться в истину на элементах множества). Переменная *rr* типизируется типом записи *rec*, обладающим необходимыми для типа коллекции атрибутами. Для каждой из соединяемых коллекций заводится по переменной (*pp* и *jj*), переменные связываются квантором всеобщности # и типизируются как принадлежащие соответствующим коллекциям: *pp: IRP & jj: JobChange*. Условия из секций *match* и *check* представляются соответствующими условиями на переменные *pp* и *jj*. Необходимые функции (например, *normName*) определяются как константы в разделе ABSTRACT_CONSTANTS конструкции *PersonFusion* и типизируются в разделе *PROPERTIES*:

```
ABSTRACT_CONSTANTS normName, …
PROPERTIES
  normName: STRING_TYPE --> STRING_TYPE
```

Функция *normName* типизируется тотальной функцией из строк в строки.

3.2. Семантика включения данных в коллекции с типами индексов. Рассмотрим отличительные особенности представления в AMN операций включения данных в коллекции с типами индексов на примере операции *insertIntoPositionsFromIRP*:

```
insert into Positions ! [id: i.cik, company: i.company ]
select [ title: normTitle(i.title) ]
from  IRP i;
```

Знак «!» после идентификатора пополняемой коллекции *Positions* означает, что пара атрибутов *cik* и *company*, следующих после него, будет использоваться в коллекции в качестве ключа. В AMN такое пополнение коллекции *Positions* осуществляется в операции *insertIntoPositionsFromIRP*:

```
insertIntoPositionsFromIRP =
SELECT …
THEN
   Positions := Positions \/
   { rr | #ii.(ii: IRP &
         rr = rec(cik: ii'cik, company: ii'company) |->
         { rr1 | #ii1.(ii1: IRP & ii1'cik = ii'cik &
               ii1'company = ii'company &
               rr1 = rec(title: normTitle(ii1'title))) }
       )
   };
   …
END;
```

Множество *Positions* пополняется при помощи операции объединения множеств «\/». Пополняющее множество, как и в правиле 3.1, образуется при помощи конструкции выделения множеств по переменной *rr*. Отличие состоит в том, что переменная *rr* типизируется типом пары. Первым элементом пары является запись *rec(cik: ii'cik, company: ii'company),* соответствующая типу ключа коллекции Positions. Вторым элементом пары является множество названий должностей *title*, извлеченных из коллекции *IRP* с условием совпадения атрибутов ключа. Элементы пары соединяются знаком «|->».

4. Семантика порядка исполнения операций. Для отслеживания фактов исполнения операций в конструкции *PersonFusion* заводится переменная *state*, отвечающая состоянию потока работ:

```
VARIABLES … state, …
INVARIANT
  state: struct(PeopleLinkCreated: BOOL,
       PositionsFromIRPInserted: BOOL,
       PositionsFromJobChangeInserted: BOOL,
       EmploymentInserted: BOOL,
       PersonInserted: BOOL
    ) & …
```

Переменная типизируется в инварианте типом структуры, атрибуты которого отражают завершение каждой из операций потока работ. Предусловием исполнения каждой из операций является завершение всех других необходимых операций. Также, завершающим действием каждой операции является установка флага завершения операции переменной *state* в значение FALSE. Для операции *insertIntoEmployment* соответствующие предусловия и присваивания выглядят в AMN следующим образом:

```
OPERATIONS
…
insertIntoEmployment =
SELECT state'PositionsFromIRPInserted = TRUE &
  state'PositionsFromJobChangeInserted = TRUE
THEN
  …
  state'EmploymentInserted := TRUE
END;
```

Тело операции образовано конструкцией SELECT [BAMN]. Это означает, что действия, указанные после ключевого слова THEN исполняются только в том случае, если предикат после ключевого слова SELECT обращается в истину. Операция *insertIntoEmployment* может быть исполнена только тогда, когда обе операции *insertIntoPositionsFromIRP* и *insertIntoPositionsFromJobChange* исполнены (т.е. флаги *state'PositionsFromIRPInserted* и *state'PositionsFromJobChangeInserted* имеют значение TRUE). После завершения операции флаг *state'EmploymentInserted* также должен быть установлен в значение TRUE.

Начальная инициализация переменной state производится в разделе INITIALISATION, флаги завершения всех операций устанавливаются в значение FALSE:

```
INITIALISATION
   state := rec(PeopleLinkCreated: FALSE,
      PositionsFromIRPInserted: FALSE,
      PositionsFromJobChangeInserted: FALSE,
      EmploymentInserted: FALSE,
      PersonInserted: FALSE)
  || …
```

4 Применение формальной семантики для языка разрешения сущностей и слияния данных для верификации потоков работ интеграции данных

Формулы, отражающие свойства потока работ интеграции данных, подлежащие верификации, добавляются в инвариант конструкции AMN, отражающей семантику потока работ интеграции данных (в рассматриваемом примере такой конструкцией является *PersonFusion*). Затем семантические спецификации помещаются в программное средство Atelier B [2], где осуществляется проверка синтаксиса спецификаций и корректности типизации переменных и термов.

Затем осуществляется автоматическая генерация теорем, отражающих сохранение инварианта при выполнении операций спецификации. Теоремы доказываются с использованием автоматических и интерактивных средств доказательства Atelier B.

В качестве примера свойства потока работ интеграции данных рассмотрим одно из возможных свойств сохранения информации при извлечении данных из исходных коллекций *IRP* и *JobChange* и формировании коллекции *Person*. Свойство формулируется в виде формулы логики предикатов и добавляется в инвариант конструкции *PersonFusion*:

```
INVARIANT … &
 (state'PersonInserted = TRUE =>
  !(nn, cc, tt).(
   (#(irp).(irp: IRP & nn = irp'name &
            cc = irp'company & tt = irp'title) or
   #(jc).(jc: JobChange & nn = jc'name &
          cc = jc'company & tt = jc'appointedAs)) =>
   #(pers).(pers: Person & pers'name = normName(nn) &
         #(ee).(ee: pers'emp & ee'company = cc &
           #(pos).(pos: ee'positions &
                   pos'title = normTitle(tt)) ) ) ) ) & …
```

В формуле утверждается, что по завершении потока работ выполняется следующее свойство: для любых наборов, состоящих из имени персоны, названия компании и должности, встречающихся совместно в записях коллекции *IRP* либо в записях коллекции *JobChange*, в коллекции *Person* найдется запись с такими же значениями имени персоны (с точностью до нормализации), названия компании и должности. Таким образом, из исходных коллекций извлекается вся возможная информация о местах работы и должностях персон.

Спецификация конструкции *PersonFusion* была помещена в средство Atelier B, была осуществлена синтаксическая проверка и проверка типов спецификации. Для каждой из операций спецификации были автоматически сгенерированы по три теоремы, в совокупности утверждающие сохранение инварианта при завершении операций.

5 Заключение

В статье рассмотрены основные принципы представления формальной семантики языка HIL, предназначенного для описания сложных потоков работ интеграции данных, включающих операции трансформации, разрешения и слияния сущностей. Рассмотренная семантика применяется для формальной верификации свойств потоков работ интеграции данных. Следующим шагом в работе будет реализация семантического отображения языка HIL в виде трансформации на высокоуровневом языке, например, ATL [3] или QVT. Это позволит использовать формальную верификацию свойств потоков работ при решении задач в различных средах интеграции [21, 22, 36].

Литература

[1] Abrial J.-R. The B-Book: Assigning Programs to Meanings. Cambridge: Cambridge University Press, 1996.

[2] Atelier B, the industrial tool to efficiently deploy the B Method. http://www.atelierb.eu/

[3] ATL - a model transformation technology. 2016. - https://eclipse.org/atl/

[4] Apache Hadoop Project. 2016. - http://hadoop.apache.org/

[5] Kevin S. Beyer, Vuk Ercegovac, Rainer Gemulla, Andrey Balmin, Mohamed Eltabakh, Carl-Christian Kanne, Fatma Ozcan, Eugene J. Shekita. Jaql: A Scripting Language for Large Scale Semistructured Data Analysis. VLDB 2011.

[6] Bleiholder J., Naumann F. Data fusion // ACM Computing Surveys (CSUR), 2009. Vol. 41. Iss. 1. Article No. 1. doi: 10.1145/1456650.1456651.

[7] Bleiholder J. Data fusion and conflict resolution in integrated information systems. — Potsdam: Hasso-Plattner-Institut, 2010. D.Sc. Diss. 184 p.

[8] D. Burdick, M. A. Hernández, H. Ho, G. Koutrika, R. Krishnamurthy, L. Popa, I. R. Stanoi, S. Vaithyanathan, and S. Das. Extracting, Linking and Integrating Data from Public Sources: A Financial Case Study. IEEE Data Eng. Bull., 34(3):60–67, 2011.

[9] Daniel Calegari, Nora Szasz. Verification of Model Transformations: A Survey of the State-of-the-Art. Electronic Notes in Theoretical Computer Science. 292: 5–25, 2013.

[10] Luna Dong X., Naumann F. Data fusion — resolving data conflicts in integration // Proc. VLDB Endowment, 2009. Vol. 2. Iss. 2. P. 1654–1655.

[11] R. Fagin, P. Kolaitis, R. Miller, and L. Popa. Data exchange: semantics and query answering. Theoretical Computer Science, 336(1):89–124, 2005.

[12] Ronald Fagin, Laura M. Haas, Mauricio Hernández, Renée J. Miller, Lucian Popa, Yannis Velegrakis. Clio: Schema Mapping Creation and Data Exchange. Conceptual Modeling: Foundations and Applications, LNCS 5600:198-236. 2009.

[13] Getoor L., Machanavajjhala A. Entity resolution for big data // KDD'13: 19th ACM SIGKDD Conference on Knowledge Discovery and Data Mining Proceedings, 2013. P. 1527–1527.

[14] IBM InfoSphere BigInsights Information Center. http://goo.gl/Bd8SE9

[15] Hernandez M., Koutrika G., Krishnamurthy R., Popa L., Wisnesky R. HIL:A high-level scripting language for entity integration // EDBT'13: 16th

Conference (International) on Extending Database Technology Proceedings, 2013. P. 549–560.

[16] Kalinichenko L.A. Method for Data Models Integration in the Common Paradigm. Proc. of the First East-European Symposium on Advances in Databases and Information Systems ADBIS'97. -- St.-Petersburg: Nevsky Dialect, 1997. -- V. 1: Regular Papers. -- P. 275--284.

[17] Kalinichenko L.A., Stupnikov S.A., Zemtsov N.A. Extensible Canonical Process Model Synthesis Applying Formal Interpretation // Avances in Databases and Information Systems: Proceedings of the East European Conference. - Berlin-Heidelberg: Springer-Verlag, 2005. - P. 183-198.

[18] Kalinichenko L.A., Stupnikov S.A. Constructing of Mappings of Heterogeneous Information Models into the Canonical Models of Integrated Information Systems // Advances in Databases and Information Systems: Proc. of the 12th East-European Conference. - Pori: Tampere University of Technology, 2008. - P. 106-122.

[19] Kalinichenko L.A., Stupnikov S.A. Heterogeneous information model unification as a pre-requisite to resource schema mapping // A. D'Atri and D. Saccà (eds.), Information Systems: People, Organizations, Institutions, and Technologies (Proc. of the V Conference of the Italian Chapter of Association for Information Systems itAIS). – Berlin-Heidelberg: Springer Physica Verlag, 2010. – P. 373-380.

[20] Kalinichenko L.A., Stupnikov S.A. OWL as Yet Another Data Model to be Integrated // Advances in Databases and Information Systems: Proc. II of the 15th East-European Conference. - Vienna: Austrian Computer Society, 2011. -- P. 178-189.

[21] Leonid Kalinichenko, Sergey Stupnikov, Alexey Vovchenko and Dmitry Kovalev. Rule-based Multi-dialect Infrastructure for Conceptual Problem Solving over Heterogeneous Distributed Information Resources // New Trends in Databases and Information Systems. Selected Papers of the 17th European Conference on Advances in Databases and Information Systems and Associated Satellite Events. – Springer, 2013. Advances in Intelligent Systems and Computing, V. 241. – P. 61-68.

[22] Kalinichenko L. A., Stupnikov S. A., Vovchenko A. E., Kovalev D. Y. Conceptual Modeling of Multi-Dialect Workflows. Информатика и ее применения, 2014. 8(4):110-124.

[23] Kopcke H., Thor A., Rahm E. Evaluation of entity resolution approaches on real-world match problems // Proc. VLDB Endowment, 2010. Vol. 3. Iss. 1-2. P. 484–493.

[24] P. G. Larsen, N. Plat, H. Toetenel. A Formal Semantics of Data Flow Diagrams. Formal Aspects of Computing 3:1993.

[25] K. Lano, J. Bicarregui, A. Evans. Structured Axiomatic Semantics for UML Models // Rigorous Object-Oriented Methods: Proc. of the Conference. – <http://www.bcs.org/upload/pdf/ewic_ro00_paper5.pdf. 2000>.

[26] K. Lano, S. Kolahdouz-Rahimi, T. Clark. Language-Independent Model Transformation Verification. Verification of Model Transformations: Proceedings of the Third International Workshop on Verification of Model Transformations. CEUR Workshop Proceedings 1325:36-45, 2014.

[27] Donald Miner. MapReduce Design Patterns: Building Effective Algorithms and Analytics for Hadoop and Other Systems. O'Reilly Media, 2012.

[28] F. Naumann, A. Bilke, J. Bleiholder, and M. Weis. Data fusion in three steps: Resolving inconsistencies at schema-, tuple-, and value-level. IEEE Data Engineering Bulletin, 29(2):21–31, 2006.

[29] C. Olston, B. Reed, U. Srivastava, R. Kumar, and A. Tomkins. Pig Latin: A Not-So-Foreign Language for Data Processing. In SIGMOD, p. 1099–1110, 2008.

[30] Schema Matching and Mapping. Bellahsene, Zohra, Bonifati, Angela, Rahm, Erhard (Eds.). Springer, 2011.

[31] Stupnikov S.A., Kalinichenko L.A., Bressan S. Interactive discovery and composition of complex Web services // Advances in Databases and Information Systems: Proc. of the 10th East European Conference. LNCS 4152. - Berlin-Heidelberg: Springer-Verlag, 2006. - P. 216-231.

[32] Vassiliadis, P., Simitsis, A., Georgantas, P., Terrovitis, M., Skiadopoulos, S.: A generic and customizable framework for the design of ETL scenarios. Information Systems 30(7), 492–525 (2005).

[33] Ступников С. А. Моделирование композиционных уточняющих спецификаций: Дис. канд. техн. наук.: 05.13.17 (Теоретические основы информатики) / ИПИ РАН, диссертационный совет Д.002.073.01. – М., 2006. – 195 с.

[34] Ступников С. А. Унификация модели данных, основанной на многомерных массивах, при интеграции неоднородных информационных ресурсов. Труды 14-й Всероссийской научной конференции «Электронные библиотеки: перспективные методы и технологии, электронные коллекции» RCDL'2012. – Переславль-Залесский: Университет города Переславля, 2012. С. 67-77. Одновременная электронная публикация в CEUR Workshop Proceedings, Vol. 934, P. 42-52. http://ceur-ws.org/Vol-934/

[35] С. А. Ступников. Отображение графовой модели данных в каноническую объектно-фреймовую

информационную модель при создании систем интеграции неоднородных информационных ресурсов // Труды 15-й Всероссийской научной конференции «Электронные библиотеки: перспективные методы и технологии, электронные коллекции» RCDL'2013. – Ярославль: Ярославский государственный университет им. П. Г. Демидова, 2013. С. 193-202. CEUR Workshop Proceedings 1108:85-94. http://ceur-ws.org/Vol-1108

[36] Ступников С. А., Вовченко А. Е. Комбинированная виртуально-материализованная среда интеграции больших неоднородных коллекций данных // Электронные библиотеки: перспективные методы и технологии, электронные коллекции (RCDL 2014): Тр. 16-й Всеросс. науч. конф. – Дубна: ОИЯИ, 2014. С. 339-348. CEUR Workshop Proceedings 1297:201-210. http://ceur-ws.org/Vol-1297/

Formal Semantics and Verification of Entity Resolution and Data Fusion Operations

Sergey Stupnikov

During all the period of development of methods and tools for data integration the issues of their formal verification were arising. Three levels of integration can be distinguished: data model integration, schema matching and integration, and proper data integration. This work proposes an approach for definition of formal semantics for high-level data integration programs. The semantics is defined using a specification language supported by formal provers. The semantics is applied for verification of structured data integration workflows. Workflow properties to be verified are presented as expressions of the specification language chosen. After that a semantic specification of the data integration workflow is verified w.r.t. required properties. A practical aim of the work is to define a basis for formal verification of data integration workflows during problem solving in various integration environments.

On Crowd Sensing Back-end

© Dmitry Namiot © Manfred Sneps-Sneppe
Lomonosov Moscow State University, AbavaNet,
Moscow, Russia
dnamiot@gmail.com, manfreds.sneps@gmail.com

Abstract

This paper is devoted to the crowd sensing applications. Crowd sensing (mobile crowd sensing in our case) is a new sensing paradigm based on the power of the crowd with the sensing capabilities of mobile devices, such as smartphones or wearable devices. This power is based on the smartphones, usually equipped with multiple sensors. So, it enables to collect local information from the individual's surrounding environment with the help of sensing features of the mobile devices. In this paper, we provide the review of the back-end systems (data stores, etc.) for mobile crowd sensing systems. The main goal of this review is to propose the software architecture for mobile crowd sensing in Smart City environment. We discuss also the deployment of cloud-back-ends in Russia.

1 Introduction

Crowd Sensing (in our case - Mobile Crowd Sensing) is a relatively new sensing paradigm, which is based on the power of the crowd mobile users (mobile devices) with the sensing capabilities [1]. It is illustrated in Figure 1.

Figure 1 Mobile Crowd Sensing [2]

So, the mobile crowd sensing is all about relying on the crowd to perform sensing tasks through their sensor-enabled devices. The background for this process is very obvious. We see the increasing popularity of smartphones (wearable devices in the nearest future), already equipped with multiple sensors. So, why do not use them for collecting the local timely knowledge from the individual's surrounding environment? In this process, we can collect various data: location data, camera information, air pollution data, etc. In other words, everything that could be done through the mobile device's sensing features. As per the latest vision, we can collect data even from the individual itself – so-called cyber-physical systems [3].

Of course, this approach presents a set of challenges. The main challenges, mentioned in the scientific papers are user participation and anonymity, data sensing quality. Most of the challenges based on the fact that humans participate in the process directly or indirectly. Obviously, the performance and usefulness of crowd sensing sensor networks depend on the crowd willingness to participate in the data collection process. The human participation raises issues regarding the privacy and security of data, as well as issues of revealing of sensitive information [4]. Of course, there are issues regarding the quality and trustworthiness of the contributed data. For example, the big question is how to detect and remove data contributed by malicious users [5]. By the definition, there is no control over the crowd sensors and hence, the system cannot control their behavior. Therefore, the overall quality of the sensor readings may deteriorate if counterfeit data is received from malicious users. Then, the obvious question is how to validate the sensing data that crowd sensors provide to the system. A commonly used approach is to validate the data depending on the trust level of the crowd sensor that reports it [6].

The collection of potentially sensitive information pertaining to individuals is an important aspect of crowd sensing. For instance, sensors readings can be used to track users movements. Such tracks can profile users and this information could be used besides our crowd sensing tasks [7]. A popular approach for preserving users privacy is the depersonalization. It could be done via, removing any user identifying attributes from the sensing data before sending it to the data store. Another approach is to use randomly generated pseudonyms when sending sensed data to the data store [8].

In our paper, we will target another challenge – data stores for mobile crowd sensing. We will present a review of tools (preferably – Open Source tools) and architectures used in crowd sensing projects.

The rest of our paper is organized as follows. In section 2, we present the common models for crowd sensing data architectures. In section 3, we will discuss crowd sensing video applications. In section 4, we

Proceedings of the XVIII International Conference «Data Analytics and Management in Data Intensive Domains» (DAMDID/RCDL'2016), Ershovo, Russia, October 11 - 14, 2016

discuss mobile back-ends. Our review has been produced as part of a research project on Smart Cities and applications for Smart Cities in Lomonosov Moscow State University. The main goal of this review is to propose the software architecture for mobile crowd sensing in Smart City environment. We note also that the architecture of the system must meet the existing restrictions in the Russian Federation, which will be discussed below.

2 The common architecture for mobile crowd sensing

What are the typical requirements for mobile crowd sensing applications? The good summary has been presented in [9], for example. Namely, the requirements are:

- Minimal intrusion on client devices. The mobile device computing overhead always must be minimized. Of course, we should cover all the stages: active state (passing data to data store) and passive state (waiting for new sensing data).
- The fast feedback and minimal delay in producing stream information. It is actually a discussable point of view. Most of the sensors are asynchronous and this fact creates own requirements to gathering data, for example [10]. But in the general – yes, data must be quickly provided.
- Openness and security.
- Complete data management workflow. The application (the platform) should support all steps the data management cycle, from collection to communication.

Due to a complexity of sensing collecting process, some models propose to use local databases for accumulating data on mobile devices and subsequent replication of them. This schema is illustrated in Figure 2 [9].

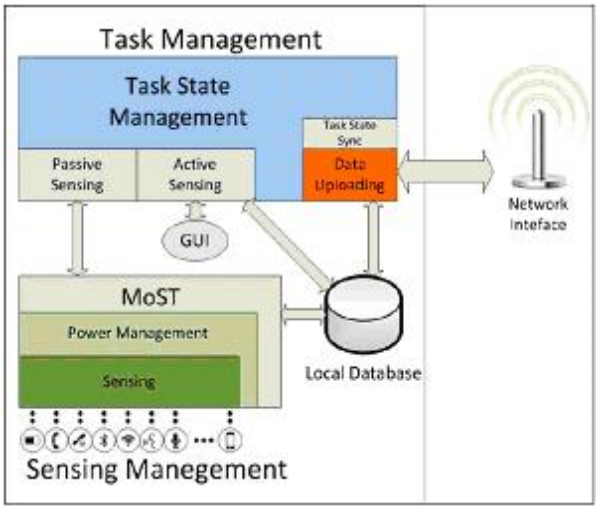


Figure 2 Local database for sensing

For example, Android platform offers several options for local data saving. The solution developers can choose depends on your specific needs, such as whether the data should be private or accessible to other applications (and users) and how much space data requires. Developers can use the following options:

- Shared Preferences. This option stores private primitive data in key-value pairs.
- Internal Storage. This option stores private data on the device memory.
- SQLite Databases. It stores structured data in a private database.

SQLite is the most often used solution here. It is a self-contained, embeddable, zero-configuration SQL database engine [11]. For example, Open Source Funf package from MIT [12] saves sensing info in SQLite database (Figure 3).

Figure 3 Funf datastore [13]

So, we can consider crowd sensing system as a set of local databases.

Another popular option in local data stores for sensing is the deployment of cloud-based file stores, like Dropbox [14]. Of course, this architecture does not assume the real-time processing, but it is simple and very easy to implement and deploy.

In the same time, many of the tasks require real-time (or near real-time) processing. In this case, the common use case is associated with some messaging bus. In this connection, we should mention so-called Lambda Architecture [15]. Originally, the Lambda Architecture is an approach to building stream processing applications on top of MapReduce and Storm or similar systems (Figure 4). Nowadays it is associated with Spark and Spark streaming too [16]. The main idea behind this schema is the fact that an immutable sequence of records is captured and fed into a batch system and a stream processing system in parallel. So, developers should implement business transformation logic twice, once in the batch system and once in the stream processing system. It is possible to combine the results from both systems at query time to produce a complete answer [17].

The Lambda Architecture targets applications built around complex asynchronous transformations that need to run with low latency. Any batch processing takes the time. In the meantime, data has been arriving and subsequent processes or services continue to work with old information. The Lambda Architecture offers a dedicated real-time layer. It solves the problem with old data processing by taking its own copy of the data, processing it quickly and stores it in a fast store. This store is more complex since it has to be constantly updated.

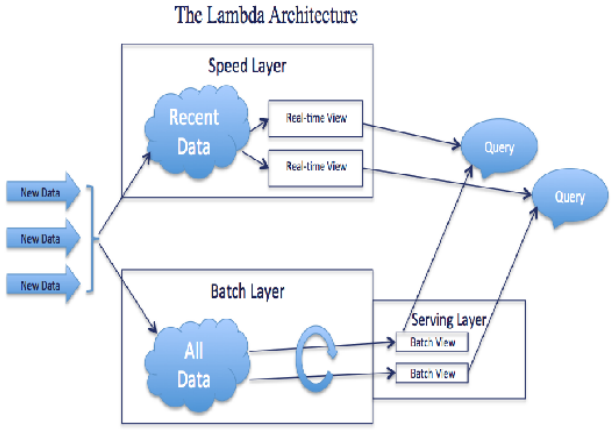


Figure 4 Lambda architecture [18]

One of the obvious disadvantages is the need for duplicating business rules. Practically, the developers need to write the same code twice – for real-time and batch layers. One proposed approach to fixing this is to have a language or framework that abstracts over both the real-time and batch framework [19].

The database (data store) design for stream processing has got own specific [20]. Broadly speaking, we have two options:

1. we can simply store every single event as it comes in (for sensing – every single measurement), dump them all in a database or a Hadoop cluster. Now, whenever we need to analyze this data in some way, we can run a query against this dataset. Of course, this will scan over essentially all the events, or at least some large subset of them;
2. we can store an aggregated summary of the measurements (events).

The big advantage of storing raw measurements data is the maximum flexibility for analysis. However, the second option also has its uses, especially when we need to make decisions or react to things in real time. Implementing some analytical methods raw data storage would be incredibly inefficient, because we would be continually re-scanning the history of measurements. The bottom line here is that raw data storage and aggregated summaries of measurements are both could be useful. They just have different use cases.

One of the prospects attempts to combine batch and real-time processing for streams is Apache Flink [21]. Flink has got a streaming dataflow engine that provides data distribution, communication, and fault tolerance for distributed computations over data streams. It is illustrated in Figure 5.

There are several Open Source solutions for data streaming support. You can find a review in our paper [15]. For example, Flume [22] is a distributed system for collecting log data from many sources, aggregating it, and writing it to HDFS. Chukwa [23] has got similar goals and features.

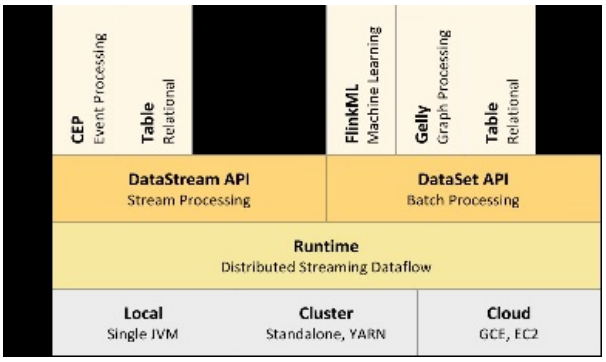


Figure 5 Apache Flink [21]

But the most used system (at least, in sensing tasks) is Apache Kafka. Apache Kafka is a distributed publish-subscribe messaging system. It is designed to provide high throughput persistent scalable messaging. Kafka allows parallel data loads into Hadoop. Its features include the use of compression to optimize performance and mirroring to improve availability, scalability. Kafka is optimized for multiple-cluster scenarios [24]. In general, publish-subscribe architecture is the most suitable approach for scalable crowd sensing applications. Technically, there are at least three possible message delivery guarantees in publish-subscribe systems:

1. At most once. It means that messages may be lost but are never redelivered.
2. At least once. It means messages are never lost but may be redelivered.
3. Exactly once. It means each message is delivered once and only once.

As per Kafka's semantics when publishing a message, developers have a notion of the message being "committed" to the log. Once a published message is committed, it will not be lost. Kafka is distributed system, so it is true as long as one broker that replicates the partition to which this message was written is still alive. In the same time, if a crowd sensing client (producer in terms of publish-subscribe systems) attempts to publish a new measurement and experiences a network error, it cannot be sure when this error happens. Is it happened before or after the message was committed? The most natural reaction for the client is to resubmit the message. It means, that we could not guarantee the message had been published exactly once. To bypass this limitation we need some sort of primary keys for inserted data. It is not easy to achieve in distributed systems. For crowd sensing systems, we can use producer's address (e.g. MAC-address or IMEI of a mobile phone) as a primary key.

Kafka guarantees at-least-once delivery by default. It also allows the user to implement at most once delivery by disabling retries on the producer and committing its offset prior to processing a batch of messages. Exactly-once delivery requires co-operation with the destination storage system (it is some sort of two-phase commit).

In connection with Kafka, we highlight two approaches. The rising popularity of Apache Spark creates the big set of projects for Kafka-Spark integration [25, 26]. And second, is the recently introduced Kafka Streams. Kafka models a stream as a log, that is, a never-ending sequence of key/value pairs. Kafka Streams is a library for building streaming applications, specifically applications that transform input Kafka topics into output Kafka topics (or calls to external services, or updates to databases, or whatever). It lets you do this with concise code in a way that is distributed and fault-tolerant [27].

On the client side for crowd sensing applications we could recommend the recently proposed by IBM Quarks System [28]. Quarks System is a programming model and runtime that can be embedded in gateways and devices. It is an open source solution for implementing and deploying edge analytics on varied data streams and devices. It can be can be used in conjunction with open source data and analytics solutions such as Apache Kafka, Spark, and Storm (Figure 6).

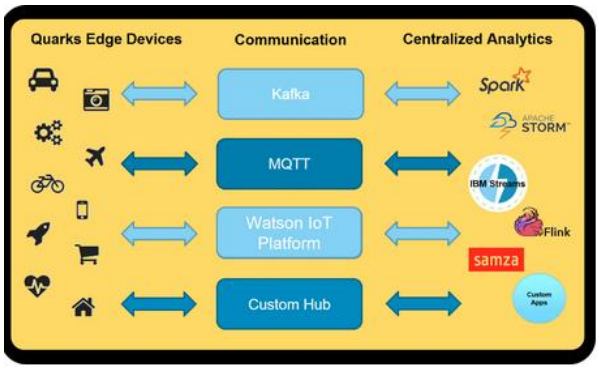


Figure 6 Quarks

As per the future, there is an interesting approach from a new Industry Specification Group (ISG) within ETSI, which has been set up by Huawei, IBM, Intel, Nokia Networks, NTT DOCOMO and Vodafone. The purpose of the ISG is to create a standardized, open environment which will allow the efficient and seamless integration of applications from vendors, service providers, and third-parties across multi-vendor Mobile-edge Computing platforms [29]. This work aims to unite the telecom and IT-cloud worlds, providing IT and cloud-computing capabilities within the Radio Access Network. Mobile Edge Computing proposes co-locating computing and storage resources at base stations of cellular networks. It is seen as a promising technique to alleviate utilization of the mobile core and to reduce latency for mobile end users [30].

We think also that 5G networks should bring changes to the crowd sensing models. In is still not clear, what is a killing application for 5G. One from the constantly mentioned approaches is so-called ubiquitous things communicating. The hope is that 5G will provide super fast and reliable data transferring approach [31]. Potentially, it could change the sensing too. 5G should be fast enough, for example, to constantly save all sensing information from any mobile device in order to use them in ambient intelligence (AMI) applications [32]. Actually, in this model crowd sensing is no more than a particular use-case for ambient intelligence. But at the moment, these are only theoretical arguments.

3 Crowd sensing for video data

In this section, we would like to discuss crowd sensing for video data. From the practical point of view, the key question here is cloud storage. Almost all existing projects use Amazon Simple Storage Service (S3) for media data (Figure 7)

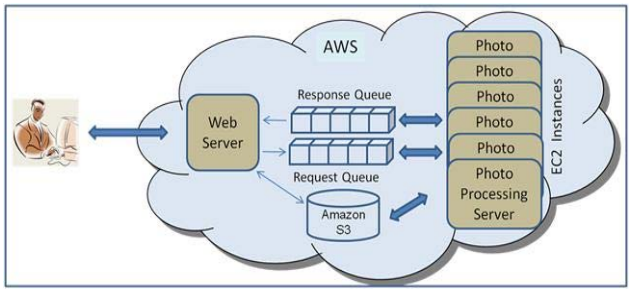


Figure 7 Amazon S3 storage

Amazon S3 is cloud storage for the Internet. It is based on the conception of buckets. To upload your data (photos, videos, documents etc.), you first create a bucket in one of the AWS regions. You can then upload any number of objects to the bucket. In terms of implementation, buckets and objects are resources, and Amazon S3 provides APIs for managing them.

Let us see, for example, the typical mobile crowd sensing application presented in [33].

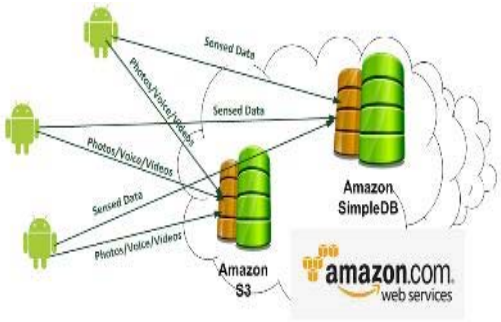


Figure 8 Amazon S3 service on practice

The cloud service provider used for this implementation is Amazon. It uses Amazon SimpleDB, a non-relational highly scalable data store. To store objects namely photos, videos and voice data, it used Amazon S3. The implementation uploads objects to S3 and maintains a key to the upload in SimpleDB. This is a basic solution. Amazon S3 stores media objects and a separate relational database (NoSQL database, e.g., key-value store) keeps keys for objects.

So, the key question here is Amazon S3 or its analogs. With the requirement to store data locally (data should not cross borders) the choice is not big. From existing Russian analogs we know about Selectel [34]. So, the real choice here is to select some Open Source platform for IaaS and build an own cloud. As Open Source platforms in this area, we can mention, for example, Cloudstack [35]. Apache CloudStack is an open source cloud computing software, which is used to build Infrastructure as a Service (IaaS) clouds by pooling computing resources. Apache CloudStack manages computing, networking as well as storage resources.

Eucalyptus (Elastic Utility Computing Architecture for Linking Your Programs To Useful Systems) [36] is free and open-source computer software for building Amazon Web Services (AWS)-compatible private and hybrid cloud computing environments.

The OpenStack project [37] is a global collaboration of developers and cloud computing technologists producing the open standard cloud computing platform for both public and private clouds.

OpenStack has a modular architecture with various code names for its components. We've mentioned just several components which are interested in the context of this paper. OpenStack Compute (Nova) is a cloud computing fabric controller, which is the main part of an IaaS system. It is designed to manage and automate pools of computer resources and can work with widely available virtualization technologies. It is an analog for Amazon EC2.

OpenStack Object Storage (Swift) is a scalable redundant storage system [38]. With Swift, objects and files are written to multiple disk drives spread throughout servers in the data center, with the OpenStack software responsible for ensuring data replication and integrity across the cluster. It lets scale storage clusters scale horizontally simply by adding new servers. Swift is responsible for replication its content.

By our opinion, the cloud solution for video in Smart City applications is a mandatory part of eco-system and OpenStack Swift is the best candidate for the platform development tool.

Note, that EU project for Smart Cities platform FIWARE proposes so-called stream generic enabler Kurento [39]. Kurento proposes public API for creating person-to-person services (e.g. video conferencing, etc.), person-to-machine services (e.g. video recording, video on demand, etc.) and machine-to-machine services (e.g. computerized video-surveillance, video-sensors, etc.). But in terms of data storage, it relies on public clouds, like Microsoft Azure.

The importance of cloud-based video services is confirmed by the industry movements. For example, we can mention IBM's newest (2016) Cloud Video Unit business [40]. As a good example (or even a prototype for the development), we can mention also Smartvue applications [41]. By our opinion, the video processing for data from moving cameras (e.g., surveillance cameras in cars) is a new hot crowd-sensing area in Smart Cities.

4 Mobile back-ends

Mobile Backend As A Service (MBaas) is a model for providing the web and mobile app developers with a way to link their applications to backend cloud storage [42]. MBaaS provides application public interfaces (APIs) and custom software development kits (SDKs) for mobile developers. Also, MBaaS provides such features as user management, push notifications, and integration with social networking services. The key moment here is the simplicity for mobile developers. As soon as many (most) of crowd-sensing applications rely on mobile phones, this direction is very interesting for crowd-sensing. Actually, the additional (to data storage) services are the key idea behind MBaaS.

As an Open Source product in this area, we can mention Convertigo [43]. It lets developers connect to enterprise data using a wide range of connectors such as SQL or Web Services, supports cross-platform development for desktop and mobile apps on multiple devices (iOS, Android), as well as server-side business logic. As another Open Source solution in this area, we can mention FIWARE cloud (Fig. 9)

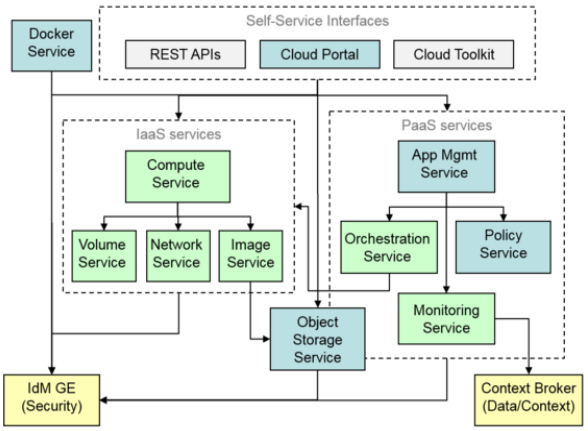


Figure 9 FIWARE mobile cloud [44]

As per [45], MBaaS offerings sit squarely between the existing platform-as-a-service vendors and the full end-to-end solution space occupied by mobile enterprise/consumer application platforms. The basic features for MBaaS include also support for programming device features (e.g., plugins or APIs) such as cameras or sensors, support for development environment (e.g., integrated version control or GIT), visual development tools, multiple operational systems support, cloud deployment, testing support, and activity monitoring. MBaaS should support user authentication (e.g., LDAP, Facebook Connect), mobile applications management, and provide task scheduler for push notifications planning [46].

5 On practical use-cases and deployment in Russia

As a conclusion for this review, we will present two use cases for back-end selection in prototype projects

with one mobile telecom operator. Firstly, it is wireless proximity information collection. Source data are network fingerprints (list of wireless nodes with signal strength). Each fingerprint has got a time stamp and could be associated (in the most cases) with some geo-coordinates. In our prototype, we use the following chain: Kafka – > Spark Streaming - > Cassandra. Cassandra has been selected as a database suitable for time series. Most of the measurements (including network proximity too) are de-facto time series data (multivariate time series). With the above mentioned chain, we can ensure the compliance with all applicable local restrictions: the personal data will be stored on the territory of the Russian Federation (all the above-mentioned components could be placed in local data centers) and Open Source components provide the absence of claims from the import-substitution point of view (this schema does not use any imported commercial software). Such a bundle is in line with modern approaches, so we can update components, reuse existing open source solutions for them, and participate in developers activities (in Open Source communities around the above-mentioned components).

The second example is much less successful. The idea of the application is data accumulation for dash cameras in vehicles. Data-saving entities are geo-coded media files (media objects). So, it is crowd-sensing for media data. The business value is transparent – city cameras cover predefined areas only, where users (cars) in the city can cover all the areas dynamically. De-facto standard for media data in a cloud is Amazon S3. But due to existing regulations (so-called personal data) it could not be used in Russia, because physically data will be saved out of the country. Alternatively, we can think about Azure Cloud Blob Storage [47] (it is a rival for Amazon S3), but there is the same question about the physical location of data outside of Russia. We think that the "standard" solution is preferable, because there are many available components and systems based on Amazon S3 API [48]. So, even the simulation of this API on the own data model lets reuse many software components. In our opinion, with the declared import-substitution and data localization regulations Amazon S3 analogue in Russia should be developed. Definitely, data centers building is not enough and we should talk about software too. It is a bit strange, why this topic is not discussed. As a base for S3 analogue development, we can probably use OpenStack (OpenStack Swift). For example, as we understand Rackspace Cloud Files is an analogue of Amazon S3 and based on OpenStack Swift.
We would like to highlight also two important moments. The above-mentioned mobile backends are oriented firstly for programming support (e.g., push notifications, social networks support, etc.). They could not solve the problems with data saving regulations, because they are oriented to existing cloud solutions (e.g., Amazon cloud).

Currently, Russia starts processes on the standardization of Internet of Things and Smart Cities. Of course, data persistence is an important part of such processes across the world and Russia could not be an exception here. Again, it looks reasonable to reuse already existing developments here. For example, we can mention such projects as oneM2M or FIWARE (there is a review for domestic standards in our paper [49]). But standards in IoT (M2M) do not provide dedicated data persistence solutions. They also rely on the existing cloud solutions. So, all the above- mentioned data saving regulations and restrictions are applicable here.

The next important trend is the strategy of vendors of sensors and other measuring devices. Many of them now include data storage as a part of "sensor". For example, Bluetooth tags Eddystone from Google include Google data storage too [50] (dislike iBeacons tags from Apple, for example). In our opinion, this trend will only rise, because data capturing lets vendors to provide additional services. It means that with the existing restrictions for data locations, the whole classes of sensors will be closed to deployment in Russia.

References

[1] Tanas, C., & Herrera-Joancomartí, J. (2013). Users as Smart Sensors: A mobile platform for sensing public transport incidents. In Citizen in Sensor Networks (pp. 81-93). Springer Berlin Heidelberg.

[2] Foremski, P., Gorawski, M., Grochla, K., & Polys, K. (2015). Energy-efficient crowdsensing of human mobility and signal levels in cellular networks. Sensors, 15(9), 22060-22088.

[3] Hu, X., Chu, T., Chan, H., & Leung, V. (2013). Vita: A crowdsensing-oriented mobile cyber-physical system. Emerging Topics in Computing, IEEE Transactions on, 1(1), 148-165.

[4] Ganti, Raghu K., Fan Ye, and Hui Lei. "Mobile crowdsensing: current state and future challenges." IEEE Communications Magazine 49.11 (2011): 32-39.

[5] Hirth, Matthias, Tobias Hoßfeld, and Phuoc Tran-Gia. "Analyzing costs and accuracy of validation mechanisms for crowdsourcing platforms." Mathematical and Computer Modelling 57.11 (2013): 2918-2932.

[6] Chen, Changlong, et al. "A robust malicious user detection scheme in cooperative spectrum sensing." Global Communications Conference (GLOBECOM), 2012 IEEE. IEEE, 2012.

[7] Beresford, Alastair R., and Frank Stajano. "Location privacy in pervasive computing." IEEE Pervasive computing 1 (2003): 46-55.

[8] Konidala, Divyan Munirathnam, et al. "Anonymous authentication of visitors for mobile crowd sensing at amusement parks." Information Security Practice and Experience. Springer Berlin Heidelberg, 2013. 174-188.

[9] Bellavista, Paolo, et al. "Scalable and Cost-Effective Assignment of Mobile Crowdsensing Tasks Based on Profiling Trends and Prediction: The ParticipAct Living Lab Experience." Sensors 15.8 (2015): 18613-18640.

[10] Namiot, Dmitry, and Manfred Sneps-Sneppe. "On software standards for smart cities: API or DPI."

ITU Kaleidoscope Academic Conference: Living in a converged world-Impossible without standards?, Proceedings of the 2014. IEEE, 2014.

[11] Yue, Kun, et al. "Research of embedded database SQLite application in intelligent remote monitoring system." Information Technology and Applications (IFITA), 2010 International Forum on. Vol. 2. IEEE, 2010.

[12] Namiot, Dmitry, and Manfred Sneps-Sneppe. "On Open Source Mobile Sensing." Internet of Things, Smart Spaces, and Next Generation Networks and Systems. Springer International Publishing, 2014. 82-94.

[13] Funf Journal http://funf.org/gettingstarted.html Retrieved: Jul, 2016

[14] Novak, Gabor, Darren Carlson, and Stan Jarzabek. "An extensible mobile sensing platform for mhealth and telemedicine applications." Proceeding of Conference on Mobile and Information Technologies in Medicine (MobileMed 2013), At Prague, Czech Republic. 2013.

[15] Namiot, Dmitry. "On Big Data Stream Processing." International Journal of Open Information Technologies 3.8 (2015): 48-51.

[16] Kroß, Johannes, et al. "Stream Processing on Demand for Lambda Architectures." Computer Performance Engineering. Springer International Publishing, 2015. 243-257.

[17] Lambda architecture http://lambda-architecture.net/ Retrieved: Jul, 2016

[18] Simplifying the (complex) Lambda architecture http://voltdb.com/blog/simplifying-complex-lambda-architecture. Retrieved: Jul, 2016

[19] Questioning the Lambda Architecture http://radar.oreilly.com/2014/07/questioning-the-lambda-architecture.htm Retrieved: Jul, 2016

[20] Gal, Zoltan, Hunor Sandor, and Bela Genge. "Information flow and complex event processing of the sensor network communication." Cognitive Infocommunications (CogInfoCom), 2015 6th IEEE International Conference on. IEEE, 2015.

[21] Apache Flink http://flink.apache.org/features.html

[22] Waga, Duncan, and Kefa Rabah. "Environmental conditions' big data management and cloud computing analytics for sustainable agriculture." World Journal of Computer Application and Technology 2.3 (2014): 73-81.

[23] Chukwa https://chukwa.apache.org/ Retrieved: Jul, 2016

[24] Garg, Nishant. Apache Kafka. Packt Publishing Ltd, 2013.

[25] Kaveh, Maziar. "ETL and Analysis of IoT data using OpenTSDB, Kafka, and Spark." (2015).

[26] Maarala, Altti Ilari, et al. "Low latency analytics for streaming traffic data with Apache Spark." Big Data (Big Data), 2015 IEEE International Conference on. IEEE, 2015.

[27] Kafka Streams http://www.confluent.io/blog/introducing-kafka-streams-stream-processing-made-simple Retrieved: Jul, 2016

[28] Quarks http://quarks-edge.github.io/ Retrieved: Jul, 2016

[29] Mobile-edge computing executing brief https://portal.etsi.org/portals/0/tbpages/mec/docs/mec%20executive%20brief%20v1%2028-09-14.pdf Retrieved: Jul, 2016

[30] Beck, Michael Till, et al. "Mobile edge computing: A taxonomy." Proc. of the Sixth International Conference on Advances in Future Internet. 2014.

[31] Osseiran, Afif, et al. "Scenarios for 5G mobile and wireless communications: the vision of the METIS project." Communications Magazine, IEEE 52.5 (2014): 26-35.

[32] Namiot, D., and M. Sneps-Sneppe. "On Hyper-local Web Pages." Distributed Computer and Communication Networks. Springer International Publishing, 2015. 11-18.

[33] Sherchan, Wanita, et al. "Using on-the-move mining for mobile crowdsensing." Mobile Data Management (MDM), 2012 IEEE 13th International Conference on. IEEE, 2012.

[34] Selectel API (Russia) https://selectel.ru/services/cloud-storage/ Retrieved: May, 2016

[35] Apache CloudStack https://cloudstack.apache.org/ Retrieved: May, 2016

[36] Kumar, Rakesh, and Sakshi Gupta. "Open source infrastructure for cloud computing platform using eucalyptus." Global Journal of Computers & Technology Vol. 1.2 (2014): 44-50.

[37] OpenStack https://www.openstack.org/ Retrieved: May, 2016

[38] Wen, Xiaolong, et al. "Comparison of open-source cloud management platforms: OpenStack and OpenNebula." Fuzzy Systems and Knowledge Discovery (FSKD), 2012 9th International Conference on. IEEE, 2012.

[39] Kurento – the stream-oriented generic enabler https://www.fiware.org/2014/07/04/kurento-the-stream-oriented-generic-enabler/ Retrieved: May, 2016

[40] IBM Cloud Video https://www.ibm.com/cloud-computing/solutions/video/ Retrieved: May, 2016

[41] Smartvue http://smartvue.com/cloud-services.html Retrieved: May, 2016

[42] Gheith, A., et al. "IBM Bluemix Mobile Cloud Services." IBM Journal of Research and Development 60.2-3 (2016): 7-1.

[43] Convertigo http://www.convertigo.com Retrieved: May, 2016

[44] FI-WARE Cloud Hosting https://forge.fiware.org/plugins/mediawiki/wiki/fiware/index.php/Cloud_Hosting_Architecture Retrieved: May, 2016

[45] Michael Facemire Mobile Backend-As-A-Service: The New Lightweight Middleware? http://blogs.forrester.com/michael_facemire/12-04-25-mobile_backend_as_a_service_the_new_lightweight_middleware Retrieved: Apr, 2016

[46] Namiot, Dmitry, and Manfred Sneps-Sneppe. "Geofence and network proximity." Internet of Things, Smart Spaces, and Next Generation Networking. Springer Berlin Heidelberg, 2013. 117-127.

[47] Calder, Brad, et al. "Windows Azure Storage: a highly available cloud storage service with strong consistency." Proceedings of the Twenty-Third ACM Symposium on Operating Systems Principles. ACM, 2011.

[48] Amazon S3 REST API http://docs.aws.amazon.com/AmazonS3/latest/API/Welcome.html Retrieved: Jul, 2016

[49] Namiot, Dmitry, and Manfred Sneps-Sneppe. "On the domestic standards for Smart Cities." International Journal of Open Information Technologies 4.7 (2016): 32-37.

[50] Namiot, Dmitry, and Manfred Sneps-Sneppe. "The Physical Web in Smart Cities." Advances in Wireless and Optical Communications (RTUWO), 2015. IEEE, 2015.

Системы анализа текстов

Text Analysis Systems

Статистическая обработка общественно-политических текстов о регионах России

© Н. Н. Абрамова

ФГУП «НИЦИ при МИД России»
Москва
NAbramova@mid.ru

Аннотация

В статье описываются методы и результаты обработки большого массива общественно-политических текстов о регионах России, собранного за 15 лет, начиная с 2001 года. Статистическая обработка осуществлялась с помощью программы анализа данных AtteStat в среде электронных таблиц Microsoft Excel. Результаты работы могут быть полезны системным администраторам, планирующим размещение информации на серверах, а также экспертам, оценивающим общественно-политическую жизнь регионов России.

1 Введение

В последние годы заметно вырос интерес к мониторингу и оценке региональной информации, представленной в СМИ. Известно более 200 систем мониторинга социальных медиа. Одна из таких систем мониторинга, анализирующая интернет-тексты по теме «Социально-политическая жизнь регионов Российской Федерации» с помощью лингвистического процессора, описывается в статье [1]. Успеха в этой области добилось также российское агентство «Смыслография», которое разрабатывает коммуникационные рейтинги регионов на основе контент-анализа с использованием материалов информационно-аналитической службы Factiva.com в Топ-100 ведущих англоязычных СМИ [2]. Каждому региону присваивается общий балл, учитывающий количество его упоминаний и долю благоприятных публикаций.

В представленной работе предлагается использовать методы математической статистики для составления рейтингов регионов по количеству публикаций в СМИ. Был обработан большой массив публикаций российских СМИ по региональной проблематике за 15 лет, каждый документ которого соотнесен с одним или несколькими регионами, что позволило выявить статистические закономерности при публикации материалов о различных регионах.

2 Формирование баз данных по регионам России

2.1 Источники для формирования баз данных

Базы данных по регионам России формируются путем отбора информации из множества информационных ресурсов:

- сообщений ведущих информационных агентств России: ИТАР-ТАСС, ИНТЕРФАКС, REGNUM, РБК, ПРАЙМ, АК&М, Россия сегодня, Росбалт, lenta.ru, vz.ru, Newsru, Полит.Ру, Утро.Ру, inopressa.ru, expert.ru;
- статей из основных центральных и московских газет и журналов: АИФ, Известия, Новые известия, Независимая газета, РБК-daily, КоммерсантЪ-Daily, Комсомольская правда, ИМ Ведомости, Российская газета, Парламентская газета, Новая газета, Советская Россия, Московская правда, Московский комсомолец, Вечерняя Москва, Огонек и т.д.;
- официальной российской информации, поступающей с сайтов российских органов государственной власти (пресс-служб президента, правительства, генеральной прокуратуры, министерств, агентств, служб, надзоров, управлений).

Объем информационных ресурсов, использованных для формирования баз данных региональной информации, составил около 8 млн. документов.

2.2 Фильтрация информации

Во избежание дублирования информации исключались источники, содержащие обзоры и дайджесты. Также не допускалось попадание в базы данных расписаний спортивных мероприятий, турнирных таблиц, результатов спортивных соревнований. Кроме того, сопровождающий базу региональной информации сотрудник ежедневно просматривает и при необходимости корректирует автоматически распределенную по регионам

Труды XVIII Международной конференции DAMDID/RCDL'2016 «Аналитика и управление данными в областях с интенсивным использованием данных», Ершово, 11-14 октября 2016

информацию. Это позволяет, в частности, более точно формировать информацию, относящуюся к региону «Москва».

2.3 Методика отбора информации

В качестве среды для функционирования баз данных используется платформа IBM Lotus Domino/Notes, располагающая встроенным языком программирования Lotus Script [4]. Специальная программа, разработанная на языке Lotus Script, отбирает информацию из информационных ресурсов, содержащих региональную информацию, с помощью запросов, составленных для каждого из 85 регионов России.

Пример запроса для региона «Калужская область» приводится ниже:

((Калужск*|Калуга|Калуге|Калуги|Калугой|Калугу|Обнинск*|Балабаново|Медын*|Малоярославец*) **AND NOT** ((Калужск* **sentence** площад*)|(Калужск* **sentence** шоссе)|(Калужск* застав*)|(метро Калужск*)|метро **sentence** "Калужская"|"м. Калужская"|Калужско-Рижск*|Синема парк на Калужской|"Калужская пл."|"на Калужской")))

В данном запросе используются следующие шаблоны и логические операторы:

* (звездочка) - замена нескольких символов в указанной позиции слова;

«» (кавычки) – точное написание фразы;

AND NOT – запрещенные слова и словосочетания;

| (или) – содержит хотя бы одно из слов;

sentence (предложение) – разделенные оператором слова находятся в одном предложении.

В поисковых запросах также могут быть использованы другие логические операторы (ранг соответствия, абзац, операторы полей, верхнего регистра и веса).

Региональная информация распределялась также по тематическим рубрикам с помощью запросов, составленных для каждой рубрики. Пример запроса для тематики «Религиозно-политические проблемы» приводится ниже:

(религиозн*|православ*|старообряд*|церков*|ислам*|фундаментали*|клерикал*|миссионер*|католи*|РПЦ|РПЦЗ|христиан*|епископ*|архиер*|Ватикан*|иуд*|пап* римск*|пресвитер*|священник*|диакон*|патриарх*|экумени*|конфесси*|администратур*|межконфессион*|внутриконфессион*|далай-лам*|конфуцианств*|сект*&!секто*|саентолог*|мормон*|мечет*|мусульман*|шариат*|тоталитарн* братств*|митрополит*|священнослужител*|служител* культ*|суфийск* орден*|католикос*|протестант*|раввин*|дацан*|лютеран*|будди*|баптист*|масон*|пятидесятник*|евангели*|монастыр*|синод*|собор*|паломни*|хамбо-лам*|курия|курией|курии|курию)

3 Методы статистической обработки

При статистической обработке использовалась программа анализа данных AtteStat, версия 13, предназначенная для работы в среде электронных таблиц Microsoft Excel 4.1 [3]. Использовались модули программного обеспечения: проверки нормальности распределения Normal Distribution Check for Excel (NDC), корреляционного анализа Correlation Analysis for Excel (CORA), кластерного анализа Cluster Analysis for Excel (CLA).

3.1 Проверка нормальности распределения

По всем годам, начиная с 2001 до 2015, было подсчитано количество поступивших в базы данных документов и определен закон распределения документов. Всего в базы данных поступило 2653060 документов, что составляет ~ 33% от общего объема использованных информационных ресурсов.

На рис.1 приведен график зависимости количества документов от года поступления (по оси x цифра 1 соответствует 2001 году, 2 – 2002 году, 15 – 2015 году).

Проверка нормальности распределения проводилась по модифицированным критериям Колмогорова, Смирнова и хи-квадрат Фишера [3]. Оказалось, что количество поступивших за год документов в базу данных по регионам (случайная величина) подчиняется нормальному закону распределения (по всем трем критериям гипотеза о нормальности не отклонялась). Это позволяет строить прогнозы о размере региональных баз данных, что имеет значение при планировании размещения информации на серверах и закупке серверов.

3.2 Корреляционный анализ

Анализировалась теснота связи (корреляционная зависимость) между количеством документов по регионам, поступившим за два года, с помощью коэффициента корреляции Фехнера. Вычисления производились по формуле

$$r_F = \frac{C - H}{C + H} \quad ,$$

где C - число пар совпадающих знаков отклонений количества документов в анализируемые годы от соответствующих средних значений, H - число пар несовпадающих знаков.

В таблице 1 приведены значения коэффициента Фехнера для некоторых сравниваемых периодов. Значения коэффициента Фехнера указывают на наличие корреляционной связи, причем теснота связи выше для пар, у которых годы идут подряд.

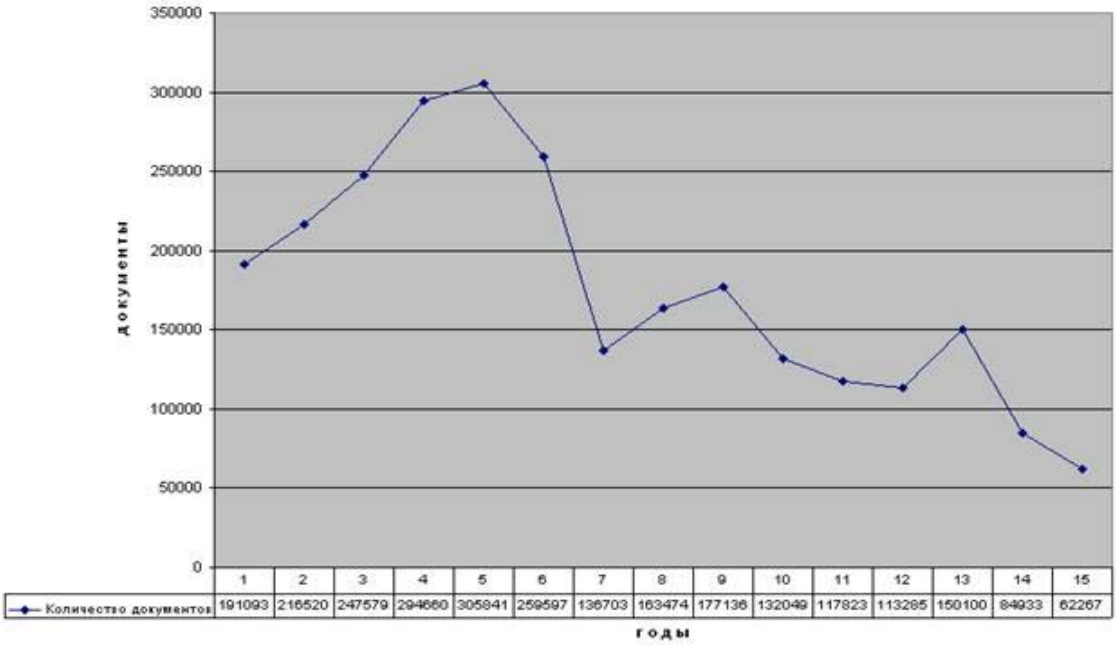


Рисунок 1 Зависимость количества документов от года поступления

Таблица 1 Корреляционная связь между количеством поступающих документов

Период (пары лет)	Коэффициент Фехнера
2001, 2015	0,6867
2002, 2003	0,9036
2002, 2013	0,7108
2004, 2005	0,8072
2008, 2009	0,8313
2009, 2010	0,9277
2012, 2013	0,8313
2013, 2014	0,8313
2014, 2015	0,7349

3.3 Кластерный анализ

Методами кластерного анализа решается задача разбиения множества регионов таким образом, чтобы все регионы, принадлежащие одному кластеру, были более похожи друг на друга по степени освещения их в прессе, чем на объекты (регионы) других кластеров. Так как оцениваются только количественные признаки, был выбран метод Уорда. Каждый объект (регион), описываемый k признаками (количество документов за один конкретный год), может быть представлен как точка в k-мерном пространстве на одном из n объектов. Сходство с другими объектами определялось как Евклидово расстояние между кластерами:

$$\rho(X_i, X_j) = \sqrt{\sum_l (x_{il} - x_{jl})^2},$$

где: X_i , X_j - координаты i-го и j-го объектов в k-мерном пространстве;
x_{il} - x_{jl} - величина l-той компоненты у i-го (j-го) объекта (l=1,2,...,k; i,j=1,2,...,n).

Сущность этого метода заключается в том, что на первом шаге каждый объект рассматривается как отдельный кластер. Процесс объединения кластеров происходит последовательно: на основании матрицы расстояний объединяются наиболее близкие объекты. Сначала объединяются два ближайших кластера. Для них определяются средние значения каждого признака и рассчитывается сумма квадратов отклонений:

$$\delta_l = \sum_i \sum_j (x_{ij} - x_{jl})^2 ,$$

где: l - номер кластера,
i - номер объекта (i = 1,2, ... , n_l),
n_l - количество объектов в l - том кластере,
j - номер признака (j = 1,2, ..., k),
k - количество признаков, характеризующих каждый объект.

В дальнейшем объединяются те объекты или кластеры, которые дают наименьшее приращение величины δ_l.

Кластеризация проводилась по 83 регионам за период с 2001 по 2013 гг. (табл. 2) и по 85 регионам (после вхождения в состав России двух новых субъектов) за 2015 гг. (табл.3).

В таблицах регионы обозначены кодами, которые используются во всех государственных органах России [5]. Например, код «77» относится к Москве, «78» - Санкт-Петербургу, «91» - Республике Крым, «92» - Севастополю и т.д.

Сравнение результатов кластеризации за 2001-2013 и 2015 годы показывает, что регионы, входящие в первую десятку, в целом, сохранили свои позиции, за исключением Чеченской республики. В группу лидеров вошли также два новых российских региона.

Можно составить рейтинг регионов по каждому году, исходя из количества документов, в которых упоминается регион. Так, в 2015 году абсолютными лидерами были Москва (1место) и С.-Петербург (2

место), затем следовали Краснодарский край, Московская область, Приморский край, Республика Крым, Севастополь, Новосибирская область, Красноярский край, и замыкал десятку Татарстан. Регионы-аутсайдеры представляли Ненецкий АО (81 место) и области: Ульяновская (82 место), Тульская (83 место), Магаданская (84 место), Липецкая (85 место). Любопытно, что Башкортостан, занявший седьмое место в рейтинге регионов в зарубежных средствах массовой информации по итогам 2015 года [2], оказался только на 16-ом месте. Такие различия можно объяснить разными подходами к отбору информации.

Таблица 2 Результаты кластеризации регионов (2001-2013 гг.)

Номер кластера	Численность кластера	Коды субъектов РФ
1	1	77
2	1	78
3	3	20,25,50
4	8	16,23,24,27,38,39,54,66
5	19	02,05,06,15,26,34,36,41, 52,53,55,59,61,63,64,65, 72,73,74
6	24	07,10,14,22,28,30,31,32, 40,42,46,47,48,49,51,56, 60,62,69,70,71,75,76,83
7	27	01,03,04,08,09,11,12, 13,17,18,19,21,29,33,35, 37,43,44,45,57,58,67,68, 79, 86,87,89

Таблица 3 Результаты кластеризации регионов за 2015 год

Номер Кластера	Численность кластера	Коды субъектов РФ
1	1	77
2	1	78
3	3	23,25,50
4	3	54,91,92
5	3	16,24,66
6	18	02,05,26,27,34,36,38,39, 41,52,55,59,61,63,65,70, 72,74
7	7	14,20,22,28,47,75,89
8	11	03,15,30,31,32,42,51,56, 60,64,69
9	22	04,06,07,10,11,17,18,19, 29,33,35,37,43,45,46,57, 58,67,68,86,87,89
10	16	01,08,09,12,13,21,40,44, 48,49,53,71,73,76,79,83

Зарубежные СМИ традиционно проявляют интерес к спортивным и статусным мероприятиям в российских регионах, таким как подготовка к чемпионату мира по футболу в 2018 году, российским спортивным первенствам, проведению саммитов ШОС и БРИКС, а в рассматриваемых нами базах данных наибольший интерес представляет общественно-политическая информация о жизни регионов.

При анализе информации о регионах важны не только количественные характеристики о числе документов, в которых упоминается регион, но и качественные, т.е. тематический характер. В качестве примера были выбраны три наиболее востребованные нашими пользователями тематики («Религиозно-политические проблемы», «Права человека», «Охрана окружающей среды» (ООС)), для которых построены графики зависимостей количества документов от года поступления (см. рис.2).

Нами была проведена кластеризация документов за 2014-2015 годы по трем перечисленным выше тематикам. В табл.4 представлены регионы, которые попали в первые четыре кластера, хотя бы по одной тематике (указывается номер кластера и в скобках рейтинг региона), а по другим тематикам они могут занимать даже последние позиции. Документы с упоминанием Москвы составили отдельные кластеры по всем тематикам, а с упоминанием Санкт-Петербурга - по двум тематикам.

4 Заключение

Проведенный анализ может быть использован для прогнозирования размера региональных баз, составления рейтинга регионов по освещению в прессе в целом и с учетом тематики, а также просмотра изменений рейтингов в динамике.

В работе показано, что наряду с методами контент-анализа, семантико-ориентированного лингвистического анализа для мониторинга публикаций о регионах можно использовать методы математической статистики – корреляционный и кластерный анализ. Эти методы не требуют затрат, связанных с работой экспертов. Однако для статистического анализа нужно располагать большим объемом информации.

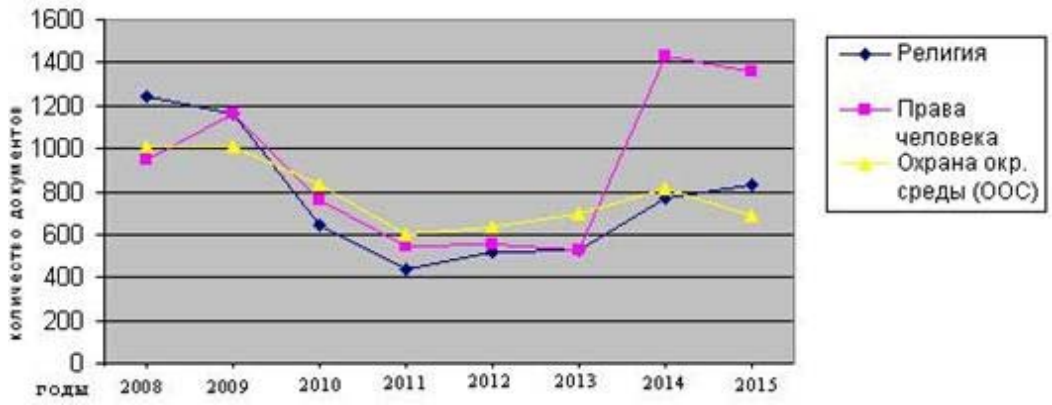


Рисунок 2 Распределение числа документов по тематикам за 2008-2013 гг.

Таблица 4 Регионы-лидеры по тематическим областям (2014 - 2015гг.)

Субъект РФ	**Религия**	**Права человека**	**ООС**
Москва	1 (1)	1 (1)	1 (1)
С.-Петербург	2 (2)	2 (2)	2 (2)
Московская обл.	3 (3)	3 (4)	2 (4)
Приморский край	5 (8)	4 (11)	2 (3)
Республика Крым	5 (9)	3 (3)	3 (7)
Севастополь	5 (10)	3 (5)	3 (9)
Краснодарский край	5 (6)	3 (6)	4 (10)
Забайкальский край	6 (25)	7 (39)	3 (5)
Красноярский край	6 (18)	5 (15)	3 (8)
Республика Бурятия	7 (45)	7 (75)	3 (6)
Новосибирская обл.	4 (5)	4 (9)	5 (17)
Чечня	4 (4)	4 (7)	7 (69)
Татарстан	4 (7)	4 (8)	6 (21)
Свердловская обл.	6 (15)	4 (12)	6 (20)
Нижегородская обл.	6 (14)	4 (10)	6 (19)
Иркутская область	6 (21)	6 (23)	4 (11)
Камчатский край	6 (23)	7 (73)	4 (12)
Хабаровский край	6 (24)	7 (40)	4 (13)

Литература

[4] Е. Б.Козеренко и др. Создание системы мониторинга интернет-текстов по теме «Социально-политическая жизнь регионов Российской Федерации»//Материалы III Международной научно-практической конференции (Москва, 18–19 сентября 2014): Сборник статей и тезисов. – М.: МГГУ им. М. А. Шолохова, 2014. С. 51–55. [Электронное издание] http://mggu-sh.ru/sites/default/files/sb_2014.pdf.

[5] О.Васильева. Коммуникационный рейтинг как инструмент оценки федерального образа региона// СОВЕТНИК №5 (185), 2011. С. 12-13 [Электронное издание] http://s-graph.ru/upload/iblock/faf/faf7de0c213d06b561daa433868163d1.jpg

[6] Программное обеспечение анализа данных AtteStat. Руководство пользователя. Версия 13//Авторское право © И.П. Гайдышев, 2002–2012. http://биостатистика.рф/files/13.pdf

[7] Н.Н.Ионцев, Е.В.Поляков, О.Г.Таранцев. Программирование в Lotus Domino R5: формулы и функции, язык Lotus Script, встоенные классы Lotus Script и Java. – М.:изд. «Светотон», 2000, 935 с.

[8] Классификатор субъектов Российской Федерации. Материалы из Википедии. https://ru.wikipedia.org/wiki/Коды_субъектов_Российской_Федерации.

Statistical processing of the social and political texts about Russian regions

N.N. Abramova

The article describes the methods and results of the processing of the social and political texts about the Russian regions that was collected since 2001 during 15 years. Statistical processing was carried out using the data analysis program AtteStat in the environment of Microsoft Excel spreadsheets. The results can be used by system administrators for planning the placement of information on the servers, as well as by experts evaluating the social and political life of the Russian Federation regions.

Тематическая классификация авторефератов диссертаций

© Ю. В. Леонова © А. М. Федотов © О. А. Федотова

Институт вычислительных технологий СО РАН,
Новосибирский государственный университет,
Государственная научно-техническая библиотека СО РАН,
Новосибирск
juli@ict.nsc.ru

Аннотация

В работе рассматривается метод тематической классификации авторефератов диссертаций. Для этого используется специально построенная мера близости документов, учитывающая специфику предметной области. Значения весовых коэффициентов в формуле для вычисления меры близости определяются предполагаемой апостериорной достоверностью данных соответствующей шкалы.

1 Предметная область

Научная тематика диссертаций разбивается на классы, называемые «специальностями». В классификаторе специальностей ВАК [1] применена трехуровневая классификация диссертаций, включающая следующие уровни: отрасль науки, группа специальностей, специальность. При этом вся система знаний разделена на 25 классов и отражает общепринятую дифференциацию наук по отраслям на физико-математические, биологические, химические, технические, сельскохозяйственные, медицинские, экономические и др. Каждый класс на лингвистическом уровне характеризуется документом, называемым паспортом специальности. Паспорт специальности устанавливает формальные критерии соответствия тематики диссертации данной специальности, в том числе определение области исследования и перечень пунктов, которым должна соответствовать диссертация.

Однако формулировки одних и тех же результатов с небольшими отличиями могут соответствовать разным специальностям. Неформальные критерии соответствия темы диссертации специальности определяются диссертационными советами, которые предъявляют разные требования к диссертациям по одной и той же специальности.

Труды XVIII Международной конференции DAMDID/RCDL'2016 «Аналитика и управление данными в областях с интенсивным использованием данных», Ершово, 11-14 октября 2016

Специальность можно характеризовать набором ключевых терминов со связями типа «родитель-потомок», представляемый в виде тезауруса. Для каждой специальности встроенный в систему общетехнический тезаурус должен дополняться своим специфическим тезаурусом, соответствующим специальности. При классификации требуется выполнить сравнение классификационных признаков автореферата, представленных ключевыми терминами с множеством ключевых терминов, представляющих паспорта специальностей, чтобы решить к какому классу относится автореферат.

2 Признаковое пространство

Признаковое пространство для классификации авторефератов представляет собой совокупность ключевых слов, терминов естественного языка и набор отношений между ними. Все выделенные отношения, которым приписаны веса, описываются в систематическом словаре понятий — тезаурусе [2]. Каждый документ описывается набором своих характеристик, называемых признаками. Определение признаков – формирование вектора характерных признаков документа, используемого в дальнейшем для принятия решений по работе с документом. Признаковое пространство представляет собой совокупность метаданных.

Формирование признакового пространства является важным этапом в задачах классификации. Необходимо отобрать наиболее значимые признаки, содержащие наибольшую информацию о классифицируемых документах. Использование слишком большого набора признаков ухудшает точность классификации, т.к. признаки содержат много избыточной информации. Признаковое пространство формируется на основе анализа текста автореферата и метаданных, содержащихся в библиографическом описании – названии, УДК, информации об оппонентах, научных руководителях и т.д. Пространство признаков разделяется на 5 типов. Каждому типу признаков сопоставляется определенный вес, указывающий его значимость [3].

1. Код УДК и шифр специальности диссертации. Вес этих признаков невысок, поскольку

проверке подвергается правильность соответствия кода УДК и шифра специальности.

2. Субъекты, которые отражены в автореферате – персоны: научные руководители, оппоненты; организации; специальность, диссертационный совет. Субъекты определяются наборами признаков – ключевых терминов, имеющих веса. Специальность определяется шифром специальности и дополнительно характеризуется ключевыми терминами из паспорта специальности. Научному руководителю ставится в соответствие набор специальностей с весами, определяемыми приоритетами научного руководителя. Для оппонента список специальностей, которые он может представлять, является равнозначным. Организации и диссертационному совету аналогично соответствует список равнозначных специальностей.
3. Ссылки на близкие работы. Выделяются персоны (субъекты), на которые ссылается автор. Персонам приписан список ключевых терминов.
4. Ключевые термины, извлекаемые из текстовых структурированных разделов автореферата: научная новизна; актуальность; цели и задачи; положения, выносимые на защиту; объект и предмет исследования; теоретическая и практическая значимость работы; методология и методы исследования; степень достоверности и апробация результатов; представление работы. Ключевые термины из разных разделов автореферата имеют разный вес. Например, практическое применение имеет низкий вес, т.к. в этом разделе не отражена тематическая направленность, а положения, выносимые на защиту – самый высокий.
5. Ключевые слова, указанные автором и экспертом. Авторские ключевые слова не всегда указываются правильно, поэтому имеют низкий вес.

Таким образом, признаковое пространство представляется в виде списков ключевых терминов и значимости. Каждому ключевому термину приписывается вектор весов (значимость) разделов, в которых встречается данный термин.

В нашем случае используется следующий порядок обработки документов. Имеется набор категорий (тем), и поступает новый документ, для которого нужно определить список подходящих ему категорий. Если документ не попадает ни в одну категорию, то он отбрасывается. На множестве категорий могут быть заданы теоретико-множественные отношения. Например, множества документов, составляющих категории, могут пересекаться или не пересекаться, т.е. один и тот же документ может принадлежать нескольким категориям. Поиск авторефератов, соответствующих определенной специальности, выполняется в направлении убывания значимости классификационных признаков. Использование классификационных признаков с весами повышает точность классификации.

3 Мера близости

Наиболее распространенным вариантом классификации документов является фасетная классификация, теория построения которой формализована индийским библиотековедом Ш.Р.Ранганатаном [4]. Объекты классифицируются одновременно по нескольким независимым друг от друга признакам (фасетам). Применительно к цифровым документам в качестве фасетов выступают элементы метаданных, в том числе и ключевые термины. Предлагаемый подход к классификации основан на понятии аналогии документов, определенной в работах [5, 6]. Ограничимся рассмотрением только ключевых терминов, агрегированных по типам признаков.

Количественная характеристика меры близости определяется на множестве документов D следующим образом [7]:

$$m: D \times D \rightarrow [0,1]$$

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-ых 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 с значением термина t_j.

Пусть $\boldsymbol{P} = \{\boldsymbol{p_k}\}_{\boldsymbol{k=1}}^{\boldsymbol{N}}$ – список уникальных ключевых терминов, входящих в оба документа, Mk – число повторений термина pk. Тогда меру близости можно переписать:

$$m(d_1, d_2) = \sum_{k=1}^{N} (a_k * M_k)(m_k / M_k),$$

где α_k – весовой коэффициент соответствующий значению термина p_k,

m_k – число совпадений термина p_k в документах d_1 и d_2.

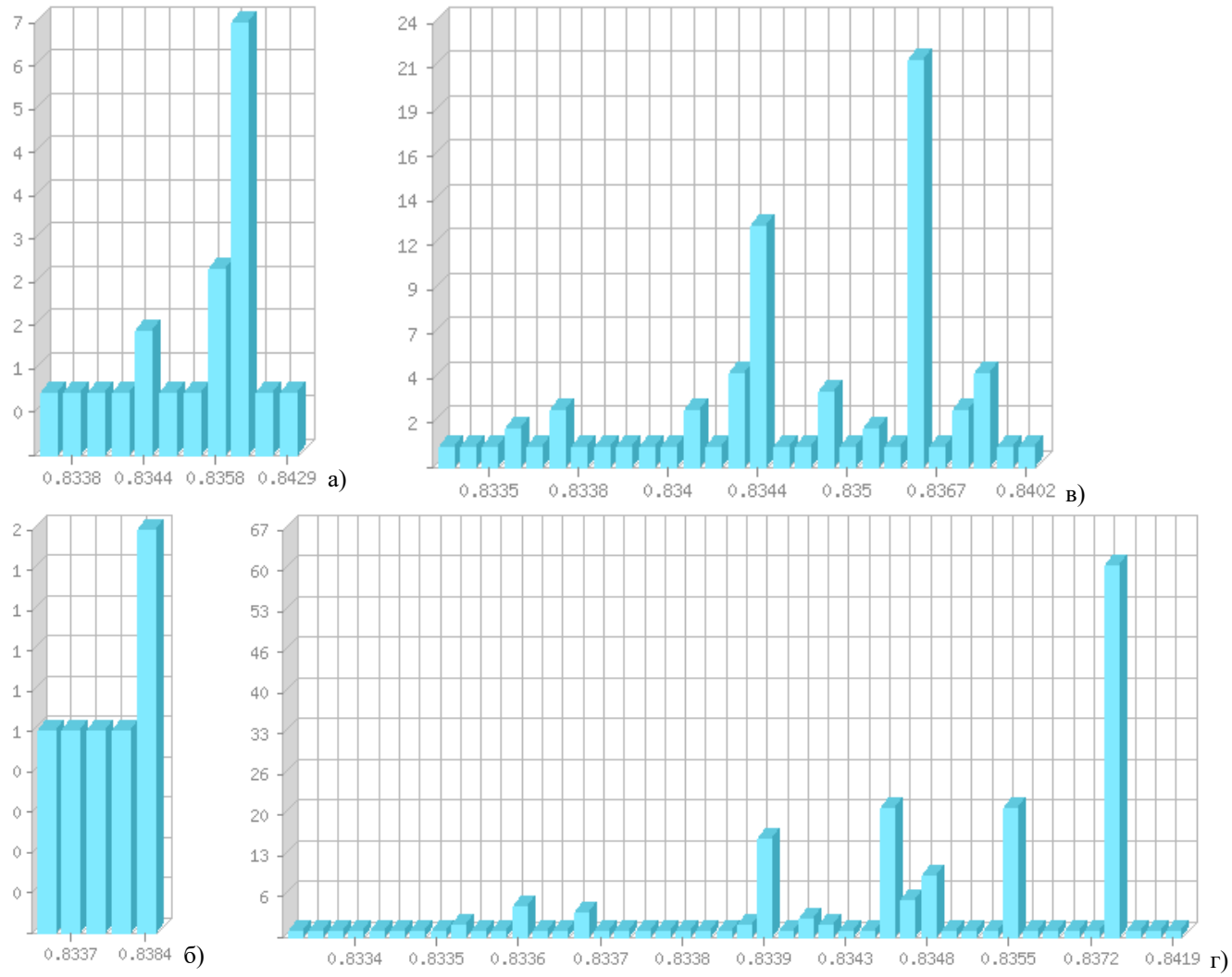


Рисунок 1 Распределение количества авторефератов по мере близости к темам а) Тема1 б) Тема2 в) Тема3 г) Тема4

Мы получаем новые весовые коэффициенты $\beta_k = \alpha_k * M_k$, которые уже характеризуют конкретный ключевой термин. Не трудно видеть, что

$$\sum_{k=1}^{N} \beta_k = 1$$

Отметим, что мы здесь автоматом получаем весовой коэффициент пропорционален частоте встречаемости термина. Кроме того, при задании меры можно принять во внимание тот факт, что значения весовых коэффициентов βk 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 документа d1 и документа d2 может быть более весомо в случае, когда количество значений этого атрибута в документе d1 достаточно велико (по сравнению со случаем, когда документ d1 имеет всего одно).

4 Тестирование алгоритма классификации. Методика

Из обучающей выборки удаляются документы рубрики, которые присутствуют при тестировании, но не участвуют в обучении.

Варианты исходов для документа:

1. «Прав»: документ («Свой») правильно определился в свою рубрику;
2. «Чуж»: действительно «Чужой» документ определился как «Чужой»;
3. «Ошиб»: документ определился не в свою рубрику;
4. «Св_чуж»: «Свой» документ ошибочно определился как «Чужой»;
5. «Чуж_св»: «Чужой» документ ошибочно попал в какую-то рубрику, т. е. ошибочно определился как «Свой».

Исходы 1 и 2 соответствуют правильной работе алгоритма, остальные – ошибочные исходы

Оценки

Точность = Прав / (Прав + Ошиб + Чуж_св)

Полнота = Прав / (Прав + Ошиб + Св_чуж)

5 Практические результаты

В качестве исходных данных для тестирования алгоритма классификации использовались авторефераты диссертаций по 3 темам: «распознавание образов» (Тема1), «распознавание речи» (Тема2), «геоинформационные системы» (Тема3), «онтологии, описание предметной области» (Тема4). В эталонные наборы для каждой тематики вошли по 30 авторефератов.

Формирование списка ключевых терминов (словаря) является отдельной задачей [7]. 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 слов.

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 вычислялось мера близости проверяемого автореферата к цетроиду класса (темы).

6 Результаты тестирования

На вход системе 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.

Принадлежность автореферата теме определяется по превышению порога близости между тестируемым авторефератом и центроидом темы.

В результате тестирования авторефераты с ненулевой мерой близости распределились по темам следующим образом: Тема1 – 76; Тема2 – 7; Тема3 – 146; Тема4 – 388.

Экспериментально установлено, что если пороговое значение меры близости превосходит 0.83, то тогда все отобранные авторефераты относятся к данной теме, при этом число не отнесённых авторефератов, посвящённых данной теме, не превосходило 5%. Если пороговое значение меры близости не превосходит 0.17, то автореферат не относится к данной теме.

Распределение количества авторефератов по уровню меры близости к темам приведены на рис. 1.

На рис.2 представлены результаты точности, полноты и F-меры, полученные при тестировании алгоритма. Наихудший параметр точности соответствует Теме3 – 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.

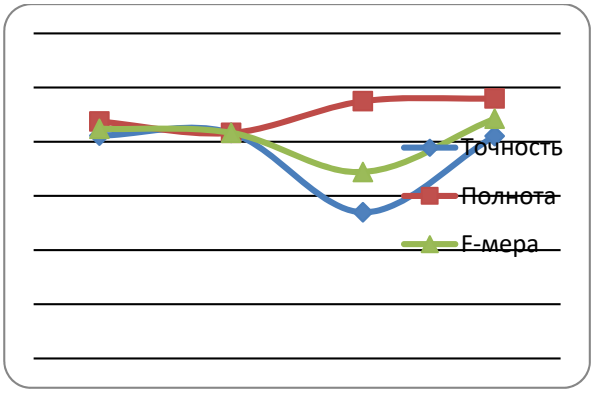


Рисунок 2 Точность, полнота и F-мера классификации авторефератов

Заключение

На основании проведенных экспериментов видно, что распределение авторефератов по мере близости к темам имеет ярко выраженный максимум при значении больше 0.83. Все документы в окрестности максимума принадлежат указанной теме. Мера близости резко отделяет значения, относящиеся к теме, от всех остальных значений, устойчиво к малым отклонениям. Таким образом, предложенный метод обеспечивает устойчивую классификацию авторефератов по темам.

Литература

[1] ОК 017-2013 Общероссийский классификатор специальностей высшей научной квалификации от 17 декабря 2013 г. N 2255-ст

[2] Гиляревский,Р.С. Рубрикатор как инструмент информационной навигации / Р.С. Гиляревский, А.В. Шапкин, В.Н. Белоозеров. – СПб.: Профессия, 2008. – 352 с.

[3] А. М. Новиков, Д. А. Новиков Методология научного исследования. – М.: Либроком, 2010, – 280 с.

[4] Ранганатан Ш.Р. Классификация двоеточием. Основная классификация / пер. с англ. М.: ГПНТБ СССР, 1970.

[5] Федотов А.М., Барахнин В.Б., Жижимов О.Л., Федотова О.А. Модель информационной системы для поддержки научно-педагогической деятельности // Вестник Новосибирского государственного университета. Серия: Информационные технологии. - 2014. - Т.12. - № 1. - С.89-101.

[6] Воронин Ю. А. Начала теории сходства. Новосибирск: Наука. Сиб. отд-ние, 1991. 128 с.

[7] Леонова Ю.В., Федотов А.М. Извлечение знаний и фактов из текстов диссертаций и авторефератов // V Международная конференция «Системный анализ и информационные технологии» (САИТ-2013): Труды конференции (Красноярск, Россия, 19-25 сентября 2013). – Красноярск: ИВМ СО РАН. – Т. 1. – 2013. – С. 232-242.

Thematic classification of theses

Yuliya V. Leonova, Anatolii M. Fedotov,
Olga A. Fedotova

In this paper, the method of thematic classification of theses is considered. It uses specially constructed measure of closeness of 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 by a posteriori reliability of the corresponding scale data.

Метод выявления заимствований в текстах разноязычных документов

© В. Н. Захаров © Ал-др А. Хорошилов © Ал-ей А. Хорошилов

ФИЦ ИУ РАН,
Москва
VZakharov@ipiran.ru Khoroshilov@mail.ru A.A.Horoshilov@mail.ru

Аннотация

В работе рассматривается метод автоматического выявления заимствований в текстах разноязычных документов, основанный на сопоставлении их формализованных представлений. При решении данной задачи была разработана модель представления смысловой структуры текстов и методы формализации и установления смысловой близости между фрагментами сравниваемых разноязычных текстов. Основным преимуществом данного метода является то, что он позволяют эффективно выявить различного рода заимствования, включая более сложные случаи плагиата.

Статья подготовлена при частичной поддержке гранта РФФИ 16-07-01028.

1 Введение

1.1 Проблема выявления заимствований в текстах документов

Наличие заимствований в работах, относящихся к сфере образования и науки, является на данный момент серьезной проблемой во многих странах мира. В связи с этим в зарубежной академической практике западных университетов и научных журналов существуют документы, регулирующие правила заимствований текста и оформления соответствующих ссылок на источники, а также четко прописаны критерии отнесения некорректных заимствований к плагиату в различных формах. Плагиатом, как правило, считается любое использование чужих идей и высказываний без должной отсылки к источнику. Заимствованием также считается пересказ текста другого источника, не сопровождающийся указанием на источник заимствования идей. В нашей стране, к сожалению, критерии выявления плагиата регламентированы не столь серьезно. Но во многих ведущих ВУЗах введены положения, которые подробно определяют ответственность учащихся за любые виды заимствований в своих работах. Для выявления заимствований во многих учреждениях образования и науки функционируют специальные информационные системы. К сожалению, возможности этих систем серьезно ограничены и они не позволяют выявлять заимствования при существенном изменении недобросовестным автором лексического состава или структуры исходного текста, а также заимствования из текстов, представленных на другом языке.

1.2 Обзор существующих подходов к задаче выявления заимствований в текстах разноязычных документов

В настоящее время задача выявления заимствований в текстах разноязычных документов недостаточно изучена в нашей стране. Поэтому не существует инструментария, позволяющего выявлять заимствования из иностранной литературы. В то же время в работах иностранных ученых эта проблема активно изучается. Так в работе [1] авторы сводят процесс поиска плагиата к трем этапам: 1) Поиск документов-кандидатов. Для этого документ автоматически переводится. Затем из документа извлекаются ключевые слова, которые после этого используются для поиска документов-кандидатов. 2) Подробный анализ документов-кандидатов. Для этого могут использоваться три поисковые модели: модель 3-грамм; явная модель семантического анализа, модель анализа подобия на основе межъязыкового выравнивания. На основе использования данных моделей принимается решение о наличии в документах-кандидатах плагиата. 3) Документы-кандидаты подробно анализируются для того, чтобы выявлять случаи, когда найденные заимствования не являются плагиатом, например, если скопированные разделы являются цитатами.

В работе [2] авторы предлагают разделить процесс поиска плагиата на 4 этапа: 1) фаза предварительной обработки (разбиение на лексемы, удаление стоп-слов); 2) извлечение ключевых слов и перевод; 3) выбор документов-кандидатов; 4) поиск

Труды XVIII Международной конференции DAMDID/RCDL'2016 «Аналитика и управление данными в областях с интенсивным использованием данных», Ершово, 11-14 октября 2016

плагиата с помощью методов, используемых для одноязычных текстов. Данный метод был разработан для сопоставления текстов на арабском и английском языках. Эксперимент показал довольно высокие показатели полноты и точности.

В работе [3] авторы предлагают метод под названием MLPlag, основанный на анализе местоположения слов. В данной работе используется тезаурус EuroWordNet для формирования независимого от языка представления текста. Детальное сравнение текстов проводится путем вычисления симметричных и асимметричных мер подобия.

Рассмотренные и другие схожие методы [1-11], разработанные зарубежными учеными, демонстрируют основные тенденции решения задачи выявления заимствований в текстах разноязычных документов. Основным недостатком, который присутствует во всех этих работах, на наш взгляд, является попытка разделять документ на отдельные слова, которые затем авторы методов пытаются перевести отдельно от контекста. Такой подход может привести к значительному числу ошибок.

2 Выявление заимствований в текстах разноязычных документов

2.1 Теоретическое представление о смысловой структуре текста

В качестве базовой теоретической концепции при разработке метода выявления заимствований в текстах разноязычных документов использовалась концепция проф. Г.Г. Белоногова и проф. Р.С. Гиляревского, констатирующая, что смысловое содержание текстов выражается с помощью единиц смысла, входящих в их состав. По их мнению, наиболее устойчивыми единицами смысла являются понятия. Проф. Г.Г. Белоногов определяет термин «понятие» как «социально значимый мыслительный образ, за которым в языке закреплено его наименование в виде отдельного слова или, значительно чаще, в виде устойчивого фразеологического словосочетания…» [14,18,27]. Понятия занимают центральное место в языке и речи и являются теми базовыми строительными блоками, на основе которых формируются смысловые единицы более высоких уровней.

Также при разработке метода были использованы конструктивные признаки текста: глобальная и локальная связности текстов [16,17,18]. Глобальная связность обеспечивает раскрытие темы документа, а локальная связность проявляется во взаимосвязи между соседними единицами текста. В соответствии с нашей моделью под глобальной смысловой связностью текста или его фрагмента будем понимать смысловую связь совокупности наименований понятий текста или его фрагмента, расположенных в определённом порядке. Под локальной смысловой связностью текста или его фрагмента будем понимать смысловую связь конкретного наименования понятия и его контекстного окружения.

Преобразование текстового представления в его формализованное смысловое представление дает возможность сопоставления текстов по их смысловому содержанию[12-13,15]. Такое сопоставление смыслового содержания текстов, обеспечивающее выявление близких по смыслу фрагментов текстов, на наш взгляд, должно удовлетворять следующим условиям:

В двух текстах должна быть пересекающаяся совокупность наименований понятий. Число понятий этой совокупности должно быть равно или превышать число наименований понятий, входящих в состав единичного высказывания.

В двух таких текстах должны быть фрагменты, в которых концентрация пересекающихся наименований понятий превышает пороговое значение. Эти фрагменты должны иметь соизмеримые размеры.

Эти фрагменты текстов должны быть сходными по составу наименований понятий и порядку их следования.

Определение схожего порядка следования наименований понятий в тексте или его фрагменте базируется на предположении, что смысл наименований понятий в значительной степени определяется их контекстным окружением [24-26]. В нашей модели смысл текста определяется как смысловое содержание совокупности взаимосвязанных наименований понятий, расположенных в нем в определенном порядке. Идентичные по смыслу тексты или их фрагменты должны удовлетворять условиям локальной и глобальной смысловой схожести. Локальная смысловая схожесть (ЛСС) наименований понятий текста определяется как сходство контекстного окружения идентичных наименований понятий в двух текстах или их фрагментах. Глобальная смысловая схожесть (ГСС) текстов или их фрагментов определяется как сходство состава идентичных наименований понятий и порядка их следования в текстах или их фрагментах. Каждое понятие этого фрагмента также должно удовлетворять условию локальной смысловой схожести.

Предлагаемая модель позволяет выявить близкие по тематике тексты или их фрагменты, после чего они, при необходимости, могут проверяться на смысловую идентичность.

2.2 Алгоритм выявления заимствований в текстах разноязычных документов

В результате проведенных исследований был разработан алгоритм выявления заимствований в текстах разноязычных документов. Необходимым условием для реализации этого алгоритма является использование многоязычного словаря унифицированных формализованных представлений наименований понятий. На данный момент в этом словаре содержатся слова и словосочетания на

русском и английском языках (общий объем словаря 3.5 млн. наименований понятий). Фрагмент многоязычного словаря унифицированных формализованных представлений наименований понятий приведен в таблице 1.

Таблица 1 Фрагмент многоязычного словаря унифицированных формализованных представлений наименований понятий

№ n/n	Основное значение в словаре	Синонимы	Эквиваленты на другом языке (английский)
…	…	…	…
816437	нефтехранилище	Нефтесклад / хранилище	oil reservoir / oil storage / petroleum storage / tank farm
816438	нефть	Каустобиолит / петролеум / черный золото	mineral oil / naphtha / oil / petrol / petroleum / rock-oil
816439	нефтяник	нефтедобытчик	Oilman / oil-industry worker
…	…	…	…

Также для работы этого алгоритма необходимы процедуры обработки текста для поддерживаемых языков. На данный момент используются процедуры для обработки текстов на русском и английском языках.

Далее приведем порядок выполнения алгоритма выявления заимствований в текстах разноязычных документов.

Шаг 1. Определяется язык анализируемого текста.

Шаг 2. Выявляется совокупность значимых наименований понятий с указанием местоположений этих понятий в тексте.

Шаг 3. Каждое наименование понятия с помощью процедуры автоматической пословной нормализации и словаря унифицированного формализованного представления наименований понятий приводится к унифицированной форме и ему присваивается номер из многоязычного словаря унифицированных формализованных представлений наименований понятий.

Шаг 4. Производится поиск совпадающих номеров наименований понятий в массиве формализованных представлений документов.

Шаг 5. Для рассматриваемого документа устанавливается перечень документов (документы могут быть на любом из поддерживаемых языков) близких ему по смысловому содержанию.

Шаг 6. Для пары документов - рассматриваемого документа и каждого из документов, найденных в п. 5, устанавливаются пары наиболее близких по смысловому содержанию фрагментов анализируемых текстов.

Шаг 7. Для каждой установленной в п.5 пары близких по смыслу фрагментов текстов определяется локальная смысловая схожесть всех наименований понятий этих фрагментов.

Шаг 8. Выбираются последовательности наименований понятий, имеющих значения локальной смысловой схожести выше заданного порога. Для каждой такой последовательности наименований понятий обоих текстов вычисляется степень их глобальной смысловой схожести.

2.3 Модель процесса выявления заимствований в текстах разноязычных документов

Модель для представления смыслового содержания текста в случае работы с разноязычными документами будет незначительно отличаться от использованной в предыдущих работах[19-23].

$КОДКО$ – концептуальный образ документа, дополненный контекстным окружением наименований понятий.

$КОДКО = \{НП_i, K_i \mid i \in [1, n_{НП}]\}$,

где $НП_i = (ННПС_i, Адр_i, \text{ОСРНП}_i, ЯНП_i)$;

$НП_i$ – информация об i-ом наименовании понятия;

$ННПС_i$ – номер наименования понятия в словаре многоязычном словаре унифицированных формализованных представлений наименований понятий;

$Адр_i$ – адреса вхождений наименования понятия в тексте;

ОСРНП_i – символ обобщенной синтаксической роли i-ого наименования понятия;

$ЯНП_i$ – язык i-ого наименования понятия;

$n_{НП}$ – количество наименований понятий;

K_i - множество контекстов i – ого наименования понятия, контексты описываются похожим образом:

$K_i = \{НПК_{ik} \mid k \in [1, n_{НПК_i}]\}$;

$НПК_{ik} = (ННПС_{ik}, Адр_{ik}, \text{ОСРНП}_{ik}, \text{КЗК}_{ik})$;

КЗК_{ik} – коэффициент значимости контекста;

Одним из важнейших этапов процесса выявления заимствований является вычисление мер выполнения условия локального и глобального смыслового сходства. Значение меры $м_{ik}$ выполнения условия локального смыслового сходства для каждого наименования понятия из КОДКО сравниваемых документов (в случае $м_{ik} = 0$ данное условие – не выполнено, при $м_{ik} > 0$ – выполнено частично, а при $м_{ik} = 1$ – выполнено полностью) вычисляется следующим образом:

Если $снп(\text{НП}_{pi}, \text{НП}_{jk}) = 0$, то $м_{ik} = 0$, иначе

$$м_{ik} = \frac{\text{снп}(\text{НП}_{pi}, \text{НП}_{jk})}{3} + \frac{2\text{ско}(\text{К}_{pil}, \text{К}_{jk\text{m}})}{3}$$

ско() – функция сравнения контекстного окружения наименований понятий;

$$\text{ско}(\text{К}_\text{a}, \text{К}_\text{b}) = \begin{cases} 1 & ,\text{фвзбк}(\text{К}_\text{a}, \text{К}_\text{b}) > 1 \\ \text{фвзбк}(\text{К}_\text{a}, \text{К}_\text{b}) & ,\text{фвзбк}(\text{К}_\text{a}, \text{К}_\text{b}) < 1 \end{cases}$$

ско() – функция вычисления значения близости контекстов;

$$\text{фвзбк}(\text{К}_\text{a}, \text{К}_\text{b}) = \frac{\sum_{c=0}^{n_{НПК_a}} \sum_{d=0}^{n_{НПК_b}} \text{фвппэ}(НПК_{ac}, НПК_{bd})}{4k_к}$$

фвппэ() – функция вычисления параметра похожести элементов контекстного окружения;

$k_к$ - размер контекста наименования понятия.

$\text{снп}(\text{НП}_{\text{pi}}, \text{НП}_{jk})$ – функция определения эквивалентности наименований понятий, причем $\text{снп}(\text{НП}_{\text{pi}}, \text{НП}_{jk}) \in \{0,1\}$, НП_{pi} – i-ый элемент формализованного смыслового описания рассматриваемого документа, НП_{jk} – k-ый элемент формализованного смыслового описания j-ого документа контрольного массива.

Условием глобального смыслового сходства является сходство порядка следования наименований понятий, но, поскольку порядок следования наименований понятий учтен при подсчете коэффициентов $м_{ik}$, с точностью до перестановок слов и словосочетаний, которые возможны в идентичных по смыслу текстах на одном языке или при переводе с одного языка на другой.

Для проверки выполнения условия глобального смыслового сходства необходимо произвести поиск последовательностей наименований понятий, у которых значения локальной смысловой схожести $м_{ik}$ выше некого заданного порога k_{ncx}. Мера выполнения условия глобального смыслового сходства вычисляется как среднее значение характеристик выполнения условия локального смыслового сходства содержащихся в этих последовательностях наименований понятий. Эта величина и будет являться коэффициентом смыслового сходства фрагментов текстов:

$$k_{cx} = \frac{\sum_{i=0}^{n_{НП_р}} \max_k(м_{ik})}{n_{НП_р}}$$

$\max_k(м_{ik})$ – максимальное значение $м_{ik}$,

при $k \in [1, n_{НП_j}]$; $n_{НП_р}$ – число элементов в КОДКО рассматриваемого документа; $n_{НП_j}$ – число элементов в КОДКО j-ого документа многоязычного контрольного массива.

3 Эксперимент выявления заимствований в текстах разноязычных документов

Для проверки работоспособности метода и возможности его использования в технологическом процессе выявления заимствований было принято решение провести небольшой эксперимент и посчитать показатели эффективности метода (полнота, точность и F1-мера). Для этого была собрана коллекция из 150 параллельных текстов (английский текст и его аутентичный перевод) по общественно-политической тематике. В процессе эксперимента русскоязычные тексты делились на предложения, для каждого из предложений определялись наиболее близкие по смысловому содержанию предложения англоязычных текстов. Пример установления смысловой близости двух разноязычных текстов приведен в таблице 2.

Таблица 2 Фрагменты параллельных текстов

Текст на русском языке	Текст на английском языке
….. Российские лидеры, конечно, беспокоятся о ценах на нефть, и для этого есть серьезная причина. Из-за падения цен на нефть падает стоимость рубля, сильно зависящая от этого показателя. Экспорт нефти важен для федерального бюджета и баланса внешней торговли России. Действительно, когда месячный курс цен на нефть марки Brent подскочил до 125 долларов за баррель в марте 2012 года, стоимость рубля приближалась к своему пику, около 29 рублей за один доллар. Когда цены на нефть упали до 30,70 доллара за баррель в январе 2016 года, стоимость рубля упала до 80 рублей за доллар. …..	….. Russia's leaders certainly do care about oil prices, and with good reason. Plunging oil prices decrease the ruble's value, which closely follows oil prices. Oil exports are important to Russia's federal budget and to its overall balance of trade. Indeed, when monthly average Brent oil prices peaked at about $125 per barrel in March 2012, the ruble was close to its own peak, at approximately twenty-nine rubles to every U.S. dollar. When Brent prices fell to $30.70 per barrel in January 2016, the ruble had fallen to about eighty rubles to the dollar. …..

Информация о текстах, участвующих в эксперименте, приведена в таблице 3.

Таблица 3 Информация о параллельных текстах

	Тексты на русском языке	Тексты на английском языке
Количество текстов	150	150
Количество предложений	6021	6021
Количество слов	157231	154863

Информация о результатах эксперимента приведена в таблице 4.

Таблица 4 Значения показателей эффективности метода

Полнота	Точность	F1 – мера
0.71	0.99	0.83

4 Заключение

В данной статье был предложен метод выявления заимствований в текстах разноязычных документов, базирующийся на семантико-синтаксическом и концептуальном анализе смысловой структуры разноязычных текстов. Разработанные на его основе алгоритмы были реализованы в виде экспериментального программного обеспечения, которое обеспечивает обработку текстов на двух языках (русском и английском). Эффективность предложенного метода была проверена на небольшой коллекции документов и показала удовлетворительные для первоначального этапа исследований результаты. Далее для улучшения качества работы метода необходимо будет провести дополнительную работу по модернизации алгоритмов и программного обеспечения, а также выполнить существенное пополнение словарей новой лексикой. Указанные мероприятия позволят значительно улучшить качество работы разработанных алгоритмов на текстах, относящихся к широкому спектру предметных областей. В настоящее время на рынке IT-услуг не существует промышленных программных средств, обеспечивающих сопоставление по их смысловому содержанию русскоязычных и англоязычных текстов. В связи с вышеизложенным нам представляется, что предлагаемый метод перспективен и кроме того он может иметь широкий спектр приложений.

Литература

[1] Potthast, Martin, Alberto Barron-Cedeno, Benno Stein, and Paolo Rosso. 2010. Cross-Language Plagiarism Detection. Language Resources and Evaluation, Special Issue on Plagiarism and Authorship Analysis, DOI: 10.1007/s10579-009-9114-z

[2] Alaa Zaid, Tiun Sabrina, Abdulameer Mohammedhasan Cross-language plagiarism of Arabic-English documents using linear logistic regression // Journal of Theoretical and Applied Information Technology, Vol. 83, No. 1, 10.01.2016, p. 20-33.

[3] Ceska Z., Toman, M, Jezek K. Multilingual Plagiarism Detection. // Artificial Intelligence: Methodology, Systems, and Applications, Proceedings of the 13th international conference on Artificial Intelligence: Methodology, Systems, and Applications, 2009, pp. 83-92.

[4] Chung-Hong Lee, Chih-Hong Wu, and Hsin-Chang Yang. 2008.A Platform Framework for Cross-lingual Text Relatedness Evaluation and Plagiarism Detection . The 3rd International Conference on Innovative Computing Information and Control (ICI-CIC'08).

[5] Mate Pataki A new approach for searching translated plagiarism . Proceedings of the 5th International Plagiarism Conference. Newcastle, UK, 2012.

[6] Ralf Steinberger Cross-lingual similarity calculation for plagiarism detection and more - Tools and resources. Keynotes for PAN 2012: Uncover-ing, Authorship, ad Social Software Misuse,2012.

[7] I.TRIFAN PLAGIARISM DETECTION IN A MULTILINGUAL ENVIRONMENT // Annals of DAAAM for 2011 & Proceedings of the 22nd International DAAAM Symposium, Volume 22, No. 1, ISSN 1726-9679, ISBN 978-3-901509-83-4, Editor B. Katalinic, Published by DAAAM International, Vienna, Austria, EU, 2011

[8] Tuomas Talvensaari Comparable Corpora in Cross-Language Information Retrieval (Academic Dissertation). Acta Electronica Universitatis Tamperensis 779, 2008.

[9] Diego Antonio Rodrıguez Torrejon, and Jose Manuel Martı Ramos Crosslingual CoReMo System. Notebook for PAN at CLEF 2011.

[10] Alexei Vinokourov, John Shawe-Taylor, and Nello Cristianini Inferring a Semantic Representation of Text via Cross-Language Correlation Analysis. Advances of Neural Information Processing Systems 15,2002.

[11] Philipp Cimiano, Antje Schultz, Sergey Sizov, Philipp Sorg, and Steffen Staab Explicit Versus Latent Concept Models for Cross-Language Information Retrieval. Proceedings of the Twenty-First International Joint Conference on Artificial Intelligence (IJCAI-09), 2009 .

[12] Кузнецов И.П. Механизмы обработки семантической информации. – М.: Наука, 1978. – 175 с.

[13] Осипов Г.С. Приобретение знаний интеллектуальными системами: Основы теории и технологии. – М.: Наука. Физматлит, 1997. – 112 с.

[14] Белоногов Г.Г. Теоретические проблемы информатики, Том 2. Семантические проблемы информатики. Под общей редакцией К.И. Курбакова. – М.: РЭА им. Г.В. Плеханова, 2008. – 342 с.

[15] Васильев В.Г., Кривенко М.П. Методы автоматизированной обработки текстов. – М.: ИПИ РАН, 2008. – 301 с.

[16] Лукашевич Н.В. Тезаурусы в задачах информационного поиска. М:, Изд. Моск. ун-та, 2011 г.- 508 с.

[17] Б. В. Добров, Н. В. Лукашевич Лингвистическая онтология по естественным наукам и

технологиям для приложений в сфере информационного поиска Учён. зап. Казан. гос. ун-та. Сер. Физ.-матем. науки, 149:2 (2007), 49–72

[18] Соссюр Фердинанд де. Курс общей лингвистики. – М.: Прогресс, 1977. – 370 с.

[19] 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. – 2012. – Вып. 8.

[20] Захаров В.Н., Хорошилов А.А. Автоматическая оценка подобия тематического содержания текстов на основе сравнения их формализованных смысловых описаний // Труды XIV-ой Всерос. науч. конф. «Электронные библио-теки: перспективные методы и технологии, электронные коллекции» – RCDL'2012, г. Переславль-Залесский, Россия, 15 – 18 октября 2012 г.

[21] Захаров В.Н., Хорошилов А.А. Методы решения задачи автоматического выявления заимствований в структурированных научно-технических документах на основе их семантического анализа // Труды XV-ой Всерос. науч. конф. «Электронные библиотеки: перспективные методы и технологии, электронные коллекции» – RCDL'2013, г. Ярославль, 14 – 17 октября 2013 года.

[22] Хорошилов А.А. Методы выявления имплицитно выраженных заимствований в научно-технических текстах на основе их концептуального анализа // Труды XVII Международ-ной конференции «Аналитика и управление данными в областях с интенсивным использованием данных» DAMDID/RCDL'2015, Обнинск, 13 – 16 октября 2015 года. С. 471-477.

[23] Хорошилов А.А. Методы, модели, алгоритмы и экспериментальное программное обеспечение автоматического выявления неявно выраженных заимствований в научно-технических текстах.: дис. … канд. техн. наук: 05.13.17: защищена 09.12.15 – М.: 2015. – 159 с.

[24] Мельчук И.А. Опыт теории лингвистических моделей «Смысл ⇔ Текст». – М.: 1974 (2-е изд., 1999).

[25] Мельчук И.А. Русский язык в модели «Смысл ⇔ Текст». – Москва – Вена, 1995.

[26] Апресян Ю.Д., Богуславский И.М., Иомдин Л.Л. и др. Лингвистическое обеспечение системы ЭТАП-2. – М.: Наука, 1989.

[27] Белоногов Г.Г., Быстров И.И. и др. Автоматический концептуальный анализ текстов. // Научно-техническая информация. Сер. 2. – М.: ВИНИТИ, 2002. – № 10.

[28] Звегинцев В.А. Предложение и его отношение к языку и речи. – М.: Изд-во Московского университета, 1976.

A method of automatic plagiarism detection in multilingual documents

Victor N. Zakharov, Alexcandr A. Khoroshilov
Alexey A. Khoroshilov

The paper presents the method of automatic plagiarism detection in multilingual documents on the base of comparison of their formalized representations. In solving this problem, we developed a model of the semantic structure of texts. To detect plagiarism, we developed an algorithm for detection of similar semantic fragments in multilingual texts. The main advantage of this method is that it makes it possible to detect not only minor changes in the structure or lexical structure of the text, but also more complicated cases in the plagiarism.

Ключевой доклад 2

Keynote Talk 2

Overview of the European Strategy in Research Infrastructures

Dimitrios Tzovaras
Information Technologies Institute,
Centre for Research and Technology Hellas,
Thessaloniki 57001, Greece
Dimitrios.Tzovaras@iti.gr

Abstract

The European Strategy Forum on Research Infrastructures (ESFRI) was established in 2002, with a mandate from the EU Council to support a coherent and strategy-led approach to policy-making on research infrastructures in Europe, and to facilitate multilateral initiatives leading to the better use and development of Research Infrastructures (RI), at EU and international level. ESFRI has recently presented its updated 2016 Roadmap which demonstrates the dynamism of the European scientific community and the commitment of Member States to develop new research infrastructures at the European level.

This work is focused on the identification of the new features and conclusions of the ESFRI Roadmap 2016 in terms of the methods and procedures that led to the call, the evaluation and selection of the new ESFRI Projects and the definition and assessment of the ESFRI Landmarks. An analysis of the impact of research infrastructures on structuring the European Research Area as well as the global research scene, and of the overall contribution to European competitiveness are also discussed focusing on data intensive RI and e-Infrastructures projects.

1 Remote Sensing Satellite Data

Remote Sensing from satellites allow a global perspective on Earth Observation to be developed. Big Data from Space is an emerging domain given the recent sharp increase of information. Fortunately, this increase is paralleled by tremendous amount of new developments related to big data in other fields and enabled by technological breakthroughs and new challenges in hardware and software developments, multi-temporal data analysis, data management and information extraction technologies. In addition, the recent multiplication of open access initiatives to big data from space is giving momentum to the field by widening substantially the spectrum of users as well as awareness among the public while offering new opportunities for scientists and value-added companies. This is especially true for Satellite data with the public release of the complete archive of Landsat data by the United States Geological Survey. At an even larger scale, the ambitious and unique European Union Copernicus Program whose Sentinel missions operated by ESA will deliver free and open access to global data in the microwave and optical/infrared ranges. The focus is on the whole data life cycle, ranging from data acquisition by space borne and ground-based sensors to data management, analysis and exploitation in the domains of Earth Observation.

We consider some recent advances to the research of Earth Observation data from satellites and evolution the EO Data Archives [1, 7]. After a long, slow rise since 1986, the volume of satellite data from ESA's satellites passed three PB and is in constant increase. New datasets, coming from the future Earth Explorer Copernicus missions will contribute to increase the volume in the years to come. Copernicus is the new name for the Global Monitoring for Environment and Security programme, previously known as GMES. Copernicus will provide information services to support policies related to environment and security at European level, but with the broader objective of operational services on a global scale. In 2015 an average of 3 TB of satellite data was generated daily. By the end of 2016 it is projected that this figure will increase to more than 6 TB a day [2]. Prediction of GSCB (Ground Segment Coordination Body) is that the volume of satellite data will exceed 20 PB for 2020.

Example 1. Russian Satellite Data Center at Vladivostok. The Pacific Center provides the studies on physics of the ocean and atmosphere and concludes reception, processing and archiving data from satellites AQUA, TERRA, MTSAT-1R, FY-1D and NOAA.

Example 2. German Aerospace Center (DLR) created the National Satellite Data Archive. The Earth Observation Center (EOC) at DLR is the Center of competence in Germany, providing expertise in Earth Observation research and development activities, as well as operational tasks for data reception, processing and archiving. The powerful and centralized archive at the DLR Earth Observation Center has proven its stability and flexibility to allow Long Term Data Preservation over more than 20 years with nearly exponentially growing data capacity. In 2012, input/output data rates have grown to be beyond 100 MB/s, but the disk drives

Proceedings of the XVIII International Conference «Data Analytics and Management in Data Intensive Domains» (DAMDID/RCDL'2016), Ershovo, Russia, October 11 - 14, 2016

and networks have also grown. Archiving capacity of National Satellite Data Archive which is Remote Sensing Data Center of the Federal Republic of Germany is 2.2 PB [3].

The ESA Earth Observation Ground Segment Department operates the Earth Observation Research and Service Support (RSS). RSS primary mission is to support the EO data user's community, to ease the development of applications adding value to raw data. The RSS environment also serves the ESA harmonization activities, collecting and classifying ground segment technology development needs. With the launch of the Sentinel-1, the RSS data farm will increase the flow rate of data by a magnitude of 100, to 1 TB of data by day.

Growing satellite data in USA. The volumes of NASA and NOAA archives have grown from 1 PB in 2000 to 10 PB in 2011 annually. Total volume of NOAA Archives will exceed 100 PB [4].

2 Spatial Data Infrastructure (SDI)

A few years ago, an idea for creating Spatial Data Infrastructure (SDI) arose. SDI is to become as an aggregate of standards for interaction of open systems, technological solutions, human resources and legal agreements for collecting, processing, distributing and using spatial data. SDI is a framework of spatial data, metadata, users and tools that connected in order to use satellite data in an efficient and flexible way. SDI is an infrastructure for spatial and digital cartographic data, which joins information resources and metadata in a GIS portal. The portal provides access to the metadata, description-based search of metadata and delivery of data. GIS portal is an integrated node of access to spatial data, which is independent of data location, format and storing structure. The development of the e-Infrastructure SDI began in the last years within the framework of INSPIRE program (Infrastructure for Spatial Information in Europe) [5]. The INSPIRE develops distributed infrastructure for geographical data for the protection of the environment in Europe, monitoring of natural resources and natural catastrophes. The geoportal INSPIRE has been created in 2005 under the European Commission initiative. The components of INSPIRE are metadata, collections and data processing services; network services and technologies; agreements on distribution, access and usage of data; mechanisms for coordinating and monitoring. INSPIRE aims at making available relevant, harmonized and quality spatial data; integrating other standards during the development process facilitates interoperability. INSPIRE infrastructure addresses technical standards, OpenGIS specifications, and protocols, including data access and the creation and maintenance of spatial information.

3 Integration of Spatial Data and Services on SDI and INSPIRE programs

The problem of data exchange is one of the most important problems in the development of SDI. This is due to difficulties in getting and processing satellite data within single infrastructure. First, this problem is related to various possibilities and requirements of providers and consumers and a lack of single standard and approach to the resources delivery. Information systems, which provide transmission on the Web of the satellite data are being intensively developed on the foundation of SDI and INSPIRE programs. The integration of spatial information and services is formed with uniform standards and protocols for data exchange. The preferred standards are ISO/TC 211 19115, CEN, OGC and W3C. The common goals of SDI and INSPIRE programs are providing coordinated distributed access to satellite informational resources; supporting solution of fundamental and applied problems related to Earth Observation from Space. The European Union's INSPIRE program is ambitious international project, which can serve as an example of implementation of the SDI, including space data.

Russian Federation began creating spatial data infrastructure for electronic exchange of spatial data and distributed access to cartographic products over the Web. The concept of *e-infrastructure* for spatial data, developed by Russian Cartography [6], defines SDI *e-Infrastructure* as a system, which provides interaction for end-users using various digital spatial data in their work. The goal of creating SDI is forming a consolidated information environment, which provides search, publication and exchange of various geographical information resources. SDI is a hierarchically arranged system build on informational technologies and based on common standards for spatial data and metadata. SDI also consists of a network of geographical information nodes — geoportals and metadata catalogues. A geoportal is a core of the informational infrastructure; it is an essential element for internal and external data exchange. A portal is a key constituent of infrastructure; it is an informational node, which contains interface for access to database (issue-related collections of satellite data, informational products of space monitoring) and a metadata catalogue.

4 Service Support Environment

Service Support Environment (SSE) is Open SOA-based environment for users and providers of data and services based on EOLI-XML, using SOAP message exchanging protocol and WSDL [7, 8]. EOLI (Earth Observation Link) is the European Space Agency's client for Earth Observation Catalogue and Ordering Services. It is used as the primary interface. EOLi works to provide browsing of the metadata and preview images of Earth Observation data acquired by the satellites: Envisat, ERS, Landsat, IKONOS, DMC, ALOS, SPOT, Kompsat, Proba, IRS, and SCISAT. SSE applies a full service oriented architecture (SOA) based on the recognized Web service standards. The system was implemented in accordance with the standards from the W3C, OASIS and the Open Geospatial Consortium organizations. Services SSE is based on different

technologies that have been integrated into a unique and powerful environment. This approach allows the service provider to offer the functionalities either as an individual Web Services or aggregated at a single interface. The main aims of creating SSE are:

- give service providers an environment that would at most simplify interaction such as "service-provider — user" and "service-provider — service-provider";
- simplify integration of existing services, providing universal XML-based interface, allowing for keeping their structure;
- provide the users with a single entry point — Internet- portal.

The core of the system consists of two main components — SSE Portal Server and AOI Server, which together form an Internet-portal, with which a user interacts. SSE Portal Server provides a web interface for users to access the portal. AOI Server works together with the SSE Portal Server and is used for giving service providers the functions of visualizing search results and of visual selection of an area on the map (Area of Interest, AOI) when specifying search criteria. Since the majority of services requires specifying geographical area as one of the search parameters, SSE Portal provides special support of this feature. When forming a query, user should specify a geographical area. User can do it in several ways:

1. choose from a list,
2. specify an area on the map,
3. upload a file, describing AOI,
4. specify an area on the map AOI by providing coordinates explicitly.

SSE Portal processes user's request using applet, loaded from AOI server, for displaying selected area on the map. SSE can visualize both locally stored information and additional map layers, downloadable from remote OGC WMS servers. The description of the selected area is stored in GML (Geography Markup Language) format and is sent to the service provider as a SOAP-message.

5 Forming the Russian Segment of SDI

The structure of SSE was initially designed based on principle "one service provider — one data array". While forming the Russian segment of SDI, another approach is used based on what has become traditional three-level information system: data level, computations level, knowledge level. Using three-level scheme is a promising way of building information environments for supporting the Earth Observation. A new feature was implemented in the scheme of accumulating and exchanging data in the Russian segment of the SDI: appearance of conceptually different components — so-called —"network nodes" — which aggregate data from several service providers without the necessity of installing and setting-up additional software. In such a connection scheme all the required data conversion to a single standard is realized in one node, while in a standard scheme this work needs to be done individually by each service provider. Besides, a user will be able to search in all the data arrays of all the service providers, which connected to the system. This system uses modern technologies, standards and open-source software, which is necessary to solve all the problems on all the development levels. These are the standard for geographical metadata of ISO series 19115-2, widely acknowledged technologies WSDL, SOAP and WMS, which are the basis for the search engine, and Eclipse development environment.

Russian segment on the SDI infrastructure constructed in IKI, which currently joins two independent satellite data providers: the information on terrains of the Pacific region supplied by IAPU at Vladivostok and satellite data for terrains of Western and Eastern Siberia supplied by ICT (Institute of Computational Technologies) at Novosibirsk. The realized solutions are well scalable and they allow for joining a large numbers of service providers of heterogeneous data. To provide users' access to satellite data, ICT creates Siberian node for collecting, storing and processing satellite data. The main functions of the node are providing telecommunications for data collecting, archiving raw data, preliminary data processing, cataloging processed data, providing long-term storing of processed data, providing access to data and topical data processing a catalogue was created based on data storing system of ICT; the catalogue is regularly updated with SPOT 4 data. The catalogue also includes archive data of Russian territory of years 1982–2002 from Landsat series satellites. A service-oriented system is created based on this catalogue; the system is realized as a basic set of applications, working in the environment of the Tomcat application server. The user interfaces are developed using PHP and JavaScript technologies. Central Authentication Server module used to provide access to the system. CAS allows for creating multi-level system of user rights and access levels. The location of the receiving complex ensures receiving data, which cover Siberia, parts of Far East and Yakutia and lands of Ural and Central Russia. It is possible to receive data from other currently active platforms. The designed software of a network node can connect data providers by using various protocols and data exchange interfaces: SOAP, ODBC, and FTP.

6 Functionality of Russian segment

Let us consider the interaction scheme of two Satellite Centers with a Russian segment node in IKI, which use different methods of data transmission. The program-hardware complex of a segment node performs the following tasks:

- direct interaction with the SSE server, processing its search queries and returning the results (message exchange with the server by XML over SOAP);
- receiving, converting and storing metadata in the local database from the satellite centers

which are part of the segment: data exchange with ICT — through HTTP gateway; data exchange with IAPU— via ODBC interface;
- Providing Web interface to a node for local search within the segment without using SSE portal.

The node uses its own DBMS, which is needed for storing metadata and local collections of each satellite center, which is connected to the node. Satellite data themselves are not stored on the node server; instead, only links to them are in the local bases, and direct access provided via these links, when necessary. The metadata in the node database is updated automatically with specified periodicity; they synchronized with local databases of each satellite center. All the necessary conversion is done in the process, since every provider stores its data in a proprietary format. Apart from the local mirror of the metadata database, each of the data providers, which is connected to the server, corresponds to its own program module, which is used for receiving and processing of the requests and for forming the response with the search results.

The results of queries processing are consolidated, converted to a given format and sent back to SSE server. The user who sent the query gets the response as a list where he/she can select necessary data using preview and then get direct links to the data.

Conclusion

The value of big data from space depends on our capacity to extract information and meaning from them. The problems of data access and data processing issues, data exchange are one of the most important problems in the development of SDI. This is due to difficulties in getting and processing satellite data within single infrastructure. Improving the accessibility of satellite data could make a very substantial contribution to environmental monitoring. IKI experience of participation in European APARSEN project confirms the interest in satellite data [9, 10]. The initial stages of our research were carried out by IKI and IAPU in the INTAS IRIS Project: Integration of Russian Satellite Data Information Resources with the Global Network of Earth Observation Information Systems. Authors have been studied integration of Russian Satellite Data with the Global Network of Earth Observation Information Systems. In this paper, we describe a distributed infrastructure for satellite data. The Russian segment of distributed informational system has been built based on EOLI-XML and SSE technologies. Technologies and principles of building distributed systems considered in this article have great potential for further development, and even today, they provide extensive means for arranging distributed heterogeneous data and services in a single global system.

References

[1] C. Reck et all. German Copernicus Data Access and Exploitation Platform. In ESA Conference on Big Data from Space BiDS'16, Spain, 2016. http://esaconferencebureau.com/custom/16M05/bids/ALL/D2_0920_Reck.pdf

[2] C. Reck et al. Behind the Science at the DLR National Satellite Data Archive. In PV 2011 Conference Ensuring Long-Term Preservation and Adding Value to Scientific and technical data. CNES. Toulouse, France.2011.

[3] Ramapriyan H.K. Development, Operation and Evolution of EOSDIS - NASA's major capability for managing Earth science data. Workshop on Repositories in Science & Technology: Preserving Access to the Record of Science. 2011. http://www.cendi.gov/presentations/11_30_11_Ramapriyan_NASA_EOSDIS.pdf

[4] The INSPIRE Directive: a brief description. https://inspire.jrc.ec.europa.eu

[5] Development of SDI for Russian Federation. https://rosreestr.ru/activity/infrastruktura-prostranstvennykh-dannykh

[6] ESA Earth Online. How to Access EO Data https://earth.esa.int/web

[7] I. V. Nedoluzhko, O. O. Korobkova. Means used for integration of catalogues in modern European EO infrastructures // Russian Digital Libraries Journal. 2012. Issue 3. http://www.elbib.ru

[8] APARSEN project - Alliance for Permanent Access to the Records of Science in Europe Network. http://www.alliancepermanentaccess.org/index.php/about-aparsen/

[9] David Giaretta. Alliance for Permanent Access. http://www.alliancepermanentaccess.org/index.php/community/conferences/apa-conferences/apa-conference-oct-2014/

Исследовательские инфраструктуры мониторинга Земли

Research Infrastructures for Earth Monitoring

Some Aspects of Development of Virtual Research Environment for Analysis of Climate Change Consequences

© Gordov E.P.[1], © Okladnikov I.G.[1], © Titov A.G.[1], © Fazliev A.Z.[2]
[1] Institute of Monitoring of Climatic and Ecological Systems of the SB of the RAS, Tomsk 634055, Russia
[2] Institute of Atmospheric Optics of the SB of the RAS, Tomsk 634055, Russia
gordov@scert.ru, oig@scert.ru, titov@scert.ru, faz@iao.ru

Abstract

We present general structure, elaborated approach and preliminary results of a project aimed at analyzing climate data and predicting impacts of climate change on the environment. One of the project objectives is provision of specialists working in climate related sciences and decision-makers with accurate and detailed climatic characteristics for the selected area and reliable and affordable tools for their in-depth statistical analysis and studies the effects of climate change. Its ultimate goal is the development of hardware and software prototype of a virtual research environment (VRE) for the climate and environmental monitoring and analysis of the impact of climate change on socio-economic processes of local and regional scale. This environment will integrate the known and new sets of climate data, software implementations of classic and new methods of statistical analysis of large data sets. It will provide the opportunity to scientists and decision-makers to use different geographically distributed spatially referenced data, processing resources and services through a web browser by integrating distributed systems that store, process and provide information via Geoportal.

The first project results, namely the scheme of large sets of geospatial climate data storage together with supporting metadata database architecture and designed an intuitive graphical user interface, are presented. Also an approach to solution of the reduction problem in quantitative climatology is discussed.

1 Introduction

Understanding of complex mechanisms of the changing climate and its effects on the environment requires generating and analyzing continuously increasing volume of observation and modeling georeferenced data [1]. Increase of diversity and volume of spatial data sets makes it impossible to collect them in a single physical archive to be processed and analyzed on the basis of traditional "working place" approaches [2]. At the same time petabyte level growth of environmental data volumes and necessity to store, search, share, transfer, process, analyze, and visualize those made the area a field for approaches and tools developed in the recently appeared data intensive application domain [3-7]. Examples of data-intensive areas are given in Ref. 6. Since volume of weather and climatic data collected and produced amounts up to petabytes they correspond to the 5V (Volume, Velocity, Variety, Variability, Veracity) model [7] and fall within the definition of "big data". To indicate the presence of geospatial dependence of physical quantities it is more correct in this case to use term "Geospatial Big Data" [8].

For comprehensive usage of large sets of georeferenced meteorological and climatic data it is necessary to create a distributed software infrastructure [9, 10], based on the spatial data infrastructure (SDI) approach [11]. SDI geoportal [12, 13] is considered as a single point that provides functionality of searching geographic information resources, retrieval of the samples according to the specified parameters (data access functionality) as well as processing and cartographical visualization services along with corresponding client applications [14]. Currently, it is generally accepted that the development of client applications as integrated elements of such infrastructure should be based on the usage of modern web and GIS technologies [15, 16, 17, 18]. According to general requirements of the INSPIRE Directive to geospatial data visualization [19], it is necessary to provide such features as data overview, image navigation, scrolling, zooming and graphical overlay as well as displaying map legends and related meta information. That is, the basic functionality of a standard GIS should be provided.

At present there are a number of information systems and services that provide similar functionality. GeoBrain Online Analysis System (GeOnAS) provides access to satellite data (NASA, USGS) via OGC services based on the GRASS GIS open source software, and has a developed web interface based on the DHTMLX library (http://dhtmlx.com/).

ncWMS service [20] is an implementation of OGC Web Map Service (WMS) for geospatial datasets represented in NetCDF format. It is actively used for data visualization within the SDI geoportals, the limiting factor being its minimal support by standard GIS.

Proceedings of the XVIII International Conference «Data Analytics and Management in Data Intensive Domains» (DAMDID/RCDL'2016), Ershovo, Russia, October 11 - 14, 2016

Unidata THREDDS (http://www.unidata.ucar.edu/software / thredds / current / tds / TDS.html) provides access to geospatial data and metadata through OPEnDAP protocols, WMS, Web Coverage Service (WCS) and others. The product also provides data subsetting functionality by using ncWMS for visualization of results.

Open distributed architecture Boundless / OpenGeo is widely used for complex geo-information applications development [21, 22]. It consists of three layers (data, application server and graphical interface), and employs the following open source software:

1. Web mapping software Geoserver + Geowebcache (http://geoserver.org), implementing OGC WMS, WFS, WPS services.
2. OpenLayers JavaScript library (http://openlayers.org/) which provides basic functionality of a thin Web GIS client.
3. GeoExt / ExtJS JavaScript library [23] for development of client web applications with an intuitively clear interface.

In this paper we outline the approach chosen for carrying out the Russian Science Foundation project [24] and specific results obtained during its first phase. In particular, we discuss the developed scheme of large sets of geospatial climate data storage and creation of relevant metadata database, as well as the designed Web GIS client with the intuitive graphical user interface. Also we describe an approach to solution of reduction problem in quantitative climatology as a basis for the planned construction of the knowledge base representing properties of geophysical data to be used in a topic-based decision support system.

2 Project approach and outlines

2.1 Objectives and approach

The proposed project is aimed at provision of specialists working in affiliated sciences focused on impact assessment, adaptation strategies, and other climate related activities and decision-makers with accurate and detailed climatic characteristics and reliable, affordable tool for their in-depth statistical analysis and studies of effects of climate change in the selected area. To reach this objective a hardware and software platform prototype forming topic based VRE for comprehensive study of ongoing and possible future climate characteristic changes and analysis of their impact on regional environment will be developed. It should provide reliable climatic information required for the study of economic, political and social consequences of global climate change at the regional level.

The large project tasks that constitute four work packages (WP) combining sets of particular thematic tasks will be solved to reach this goal.

2.2 WP 1. "Preparation of geo-referenced data sets"

WP 1 is dedicated to the creation of the data archive reflecting the detailed picture of the ongoing climatic processes and their future projections for the period 1960-2100 years and its integration into VRE under development. The archive should comprise detailed data on climatic characteristics of the study area, detailed description of the surface and its changes and form a solid foundation required to study the response of economic, social and political processes on on-going and future climatic processes.

2.3 WP 2 "Improvement of methods of analysis of climate change"

WP 2 is dedicated to the development of new methods of statistical analysis of climatic processes and their implementation into computational VRE research and analysis tools. This task should provide the developed VRE with functionality to perform modern statistical analysis of spatio-temporal climatic characteristics including their extreme manifestations.

2.4 WP 3 “Functionality of the VRE prototype”

WP 3 will be devoted to enhancing the functionality of the VRE prototype in order to create the intuitively clear and reliable tools for investigations of regional social, economic and political consequences of climate change. The result of this task will be development of opportunities for experts in the field of environmental sciences to carry out the study of the impact of climate change on relevant processes in the study region without getting a second education in computer science and their applications.

2.5 WP 4. "Case study of Western Siberia and dissemination of project results"

This task should demonstrate the VRE potential to its end user, namely to local stakeholders, decision makers and experts in the field of economic, social and political sciences. In particular, it will be shown that the elaborated VRE prototype allows one to study the climate change impact on processes in the region without mastering modern statistical analysis of geo-referenced data and advanced programming. To this end a topical DSS using ontology knowledge bases will be developed.

Specific results of this work package are detailed maps of different characteristics of extreme events and trends (with estimates of statistical significance), including estimates of return periods for certain disasters and uncertainty estimates for different types of extreme events. These maps will be available in the VRE providing easy access to interested consumers.

3 The first results

Part of tasks of the work packages WP2 and WP4 has been performed on the first stage of the project carrying out. Approaches used to solve related problems and results obtained are described in this section.

3.1 Metadata database

3.1.1 Data storage scheme

Currently, two major approaches to storing geospatial data are used: geospatial databases and file collections. The database approach utilizes relational and non-relational, local or distributed spatial databases such as

Apache HBase, Esri Geodatabase, Paradigm4, SciDB, etc. This approach requires inserting all data into spatial database before their actual use, which is quite disk space- and time-consuming operation. The file collections approach relies on storing data as a collection of data files in file system directories. Usually, self-describing formats are used for storing geospatial data. It was shown [25] that retrieval of data chunks larger than 40 Mb from a spatial database is less effective than from a simple data files collection. On the other hand, file-based approach provides fast data extraction in most cases, without a redundant preprocessing. However, this approach requires development of additional software layer to provide API adapters for storing and processing of distributed file collections. In our case simplicity and flexibility of file-based data storage approach prevailed over the other one.

Network Common Data Form (netCDF) was chosen as a major file format for most geospatial data in our data archive. This format is formally acknowledged by scientific institutions (including UCAR) and OGC as a standards' candidate for storing geospatial data and stimulating data exchange. Data are stored on data storage systems as collections of netCDF-files and arranged in a strict hierarchy of directories:

- /<data root directory>/
- <data collection name>/
- <spatial domain resolution>/
- <time domain resolution>/
- <files and directories with data>

Here, <*data root directory*> is a root location of data collections, <*data collection name*> is a name of the directory containing a single data collection, <*spatial domain resolution*> is a name of the directory containing data with the same horizontal resolution, <*time domain resolution*> is a name of the directory containing data with the same time step. All data files (sometimes grouped in subdirectories) are located deeper in the hierarchy. Names of files and subdirectories are not regulated and determined by the individual specifics of a particular data set. Every data file contains one or more multi-dimensional arrays of meteorological parameters.

3.1.2 Architecture of metadata database

To describe geospatial datasets and their processing routines, and provide effective VRE functioning a dedicated metadata database (MDDB) is required. Currently, there is a lack of such database in the area of Earth sciences. The first attempt of comprehensive description of climatic geospatial datasets and processing routines in a single database is characterized by the following features.

This database contains spatial and temporal characteristics of available geospatial datasets, their locations, and run options of software components for data analysis. Here the following terminology is used. "Dataset" is a set of data which is a) given on a single temporal and spatial grid, b) covers the same time range and c) obtained under the same simulation or observation conditions (if applicable). It is represented by a collection of netCDF files containing the same set of meteorological parameters. It is necessary to distinguish the term "parameter" ("meteorological parameter") and "variable". Meteorological parameter is the name of some meteorological characteristic: temperature, pressure, humidity, etc. Variable is a unique name of a multidimensional array in a netCDF. Names of meteorological parameters are standardized. In contrast, the names of variables in different datasets could be different, and usually depend on preferences of an institution which produced them. Along with data, netCDF files contain horizontal, vertical and time domain grids.

"Data collection" is a collection of datasets created by the organization within the specific project, but specified on different spatial and/or temporal grids, or for different scenarios. The collection may consist of one dataset.

Tables in MDDB are divided into "technical" and "interface". Technical tables contain data intended for computing software components. Interface tables hold string multilingual content for the graphical user interface.

There are two major parts of MDDB providing description of climate datasets and description of data processing software components (computing modules).

Each climate dataset in MDDB is uniquely identified by its four major characteristics: name of the data collection, resolution of the horizontal grid, resolution of the time grid and name of the modeling scenario (if applicable). And each dataset includes one or several data arrays containing values of various meteorological parameters given on spatial and temporal grids. Information about all available for analysis datasets is stored in the first part of MDDB. It is used to locate data files and to provide metadata on request.

Geospatial data processing is performed by a set of dedicated computing modules. These modules are run in accordance to a pipelined call sequence. This sequence is prepared by the web portal on the basis of user interactions with graphical user interface (GUI). Second part of MDDB contains description of various call sequences and their options. Since some data analysis routines are designed to process only specific meteorological parameters, connections between computing modules and data arrays are set in MDDB.

3.2 Web-GIS client

Boundless/OpenGeo architecture was used as a basis for Web-GIS client development.

A cartographical web application (Web-GIS client) for working with archive of geospatial NetCDF datasets contains 3 basic tiers [26]:

- Tier of NetCDF metadata in JSON format
- Middleware tier of JavaScript objects implementing methods to work with:
  - NetCDF metadata
  - XML file of selected calculations configuration (XML task)
  - WMS/WFS cartographical services

- Graphical user interface tier representing JavaScript objects realizing general application business logic.

3.2.1 NetCDF metadata tier

Web-GIS client metadata tier represents a set of interconnected JSON objects, created on the base of MySQL metadata relations, and presenting NetCDF datasets information (spatial and temporal resolutions, meteorological parameters available, acceptable processing procedures, etc.). Generally, there are two kinds of objects:

1. Objects with the structure conforming to the corresponding metadata database relations, for instance, object of measurement units.
2. Objects based on complex SQL queries to metadata relation sets that allow fast retrieving of necessary information using MySQL indices as associative array keys.

The structure of JSON objects was chosen according to the following criteria:

1. Efficiency of filling out graphical user interface interactive forms;
2. Optimization of process of creating and editing of XML file of selected calculations configuration (XML task).

It might be concluded that by virtue of the approach chosen the processes of interaction between user and metadata database via Web-GIS graphical interface are optimized.

3.2.2 Middleware tier of JavaScript objects

This tier implements methods to work with NetCDF metadata, XML task file and WMS/WFS cartographical services, and appears to be a middleware which connects JSON metadata and graphical user interface tiers. The methods include such procedures as:

1. Loading and updating of metadata JSON objects using AJAX technology
2. Creating, editing, serialization of XML calculation task object
3. Launching and tracking the task execution process located on the remote calculation node
4. Working with WMS/WFS cartographical services: obtaining the list of available layers, presenting layers on the map, export layers into various formats according to user request, obtaining and presenting the layer legend with the selected SLD style applied.

3.2.3 Graphical user interface

The tier is based on the conjunction of JavaScript libraries such as OpenLayers, GeoExt and ExtJS and represents a set of software components either standalone (information panels, buttons, list of layers, etc.), or implementing general application business logic (menu, toolbars, wizards, mouse and keyboard event handlers, and so on). Graphical interface performs two main functions: providing functional capabilities for editing XML task file, and visual presentation of cartographical information for the end user. It is similar to the interfaces of such popular classic GIS applications as uDig, QuantumGIS, etc. The basic elements of the graphical user interface include (Fig. 1):

1. Panel displaying user cartographical layers on the map. Google maps are used as a base layer by default, but there's a possibility to set an arbitrary base layer including newly created by the user
2. Layer tree allowing to toggle layer display
3. Layer legends display panel
4. Map information panel (scaling, cartographical projections, cursor geographical coordinates)
5. Application general status panel
6. Overview map panel
7. General application menu
8. Toolbar (adding/removing layer, saving NetCDF data, panning, map refresh, obtaining of information related to given geographical point, etc.)
9. Application context menu
10. Wizard creating cartographical layers based on results of computational processing of geospatial datasets available to the system.

The toolbar, application and context menus contain mouse and keyboard event handlers which uniquely define Web-GIS behavior depending on user actions with the execution context applied. Web-GIS client complies with the general INSPIRE standard requirements and provides computational processing services launching to support solving tasks in the area of environmental monitoring, as well as presenting calculation results in the form of WMS/WFS cartographical layers in raster (PNG, JPG, GeoTIFF), vector (KML, GML, Shape), and binary (NetCDF) formats.

It should be noted that geospatial data cartographical services based on Geoserver software can be used in Web-GIS client considered as well as in standard desktop GIS applications.

3.3 An approach to solution of reduction problem in quantitative climatology

Collected, analyzed, consistent and systematized data of subject domains related to problems demanding application of decision support systems have to be represented both on terminological and conceptual levels. In our project the representation will be implemented in form of two groups of OWL-ontologies related to different subject domains using semantic web technologies. The first group contains ontologies which represent properties of geophysical (including climatological) data. The second group contains OWL-ontologies characterizing a few simple tasks of economic and social domains demanding DSS usage. Different approaches to solution of reduction problem in implementation of quantitative climatology ontologies are considered now.

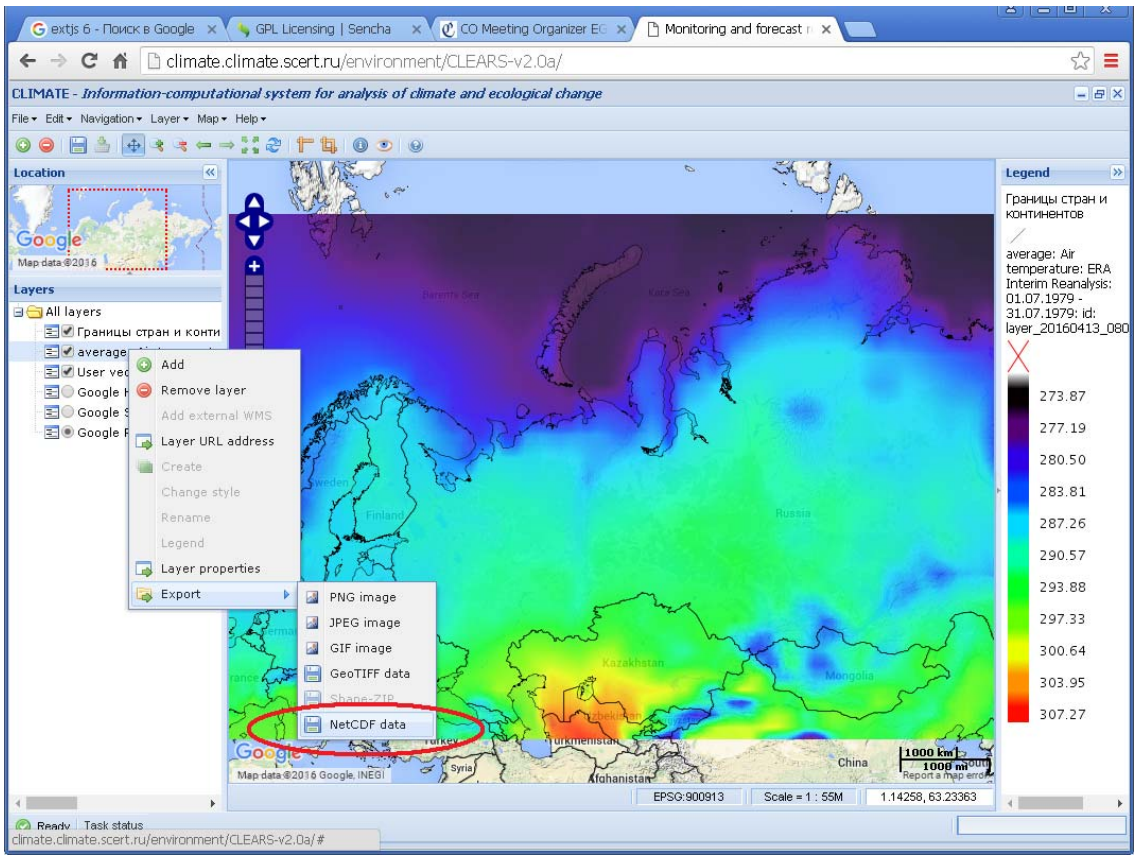


Figure 1 Graphical user interface of the web-GIS client. Exporting layer into NetCDF format is shown

4 Conclusion

To date, there is no formalized description of the metadata database for large sets of geospatial meteorological and climatic data. The architecture presented above is the first attempt to address this crucial for data intensive domain problem. Designed metadata database solves three main tasks: 1) provides content for the graphical user interface; 2) provides to geoportal information needed to generate the correct task file for the computational core; 3) contains information on the structure and arrangement of the data sets allowing computational core to read and process them efficiently. The use of this database organizes information on available data sets, facilitates the automatic retrieval of data files and improves the scalability and flexibility of computations.

The developed GIS Web client is based on the architecture Boundless / OpenGeo. The first version of the GUI uses connected JavaScript libraries, OpenLayers, GeoExt and ExtJS and is a set of software components including independent components (dashboards, buttons, layer lists), and those implementing the general logic of the application implementation (menus, toolbars, wizards, mouse and keyboard events handlers, etc.).

First application of the developed metadata database and user interface showed that their combined usage facilitates expanding of a set of data archives available for analysis and adding new statistical processing procedures [27].

Results obtained show that the developed VRE and tools would be useful for decision makers and specialists working in affiliated sciences, with the focus of their work as socio-economic impact assessment, ecological impact assessment, adaptation strategies, science policy administration and other climate related activities. On this basis they will get reliable climate related characteristics required for studies of economic, political and social consequences of global climate change at the regional level.

Acknowledgements

The authors thank the Russian Science Foundation for the support of this work under the grants 16-19-10257, 16-07-01028.

References

[1] Lykosov V.N., Glazunov A.V., Kulyamin D.V., Mortikov E.V., Stepanenko V.M. Supercomputing Modeling in Physics of Climatic System. Moscow State University Publishing House, 2012, 402 p.

[2] Gordov E.P., Kabanov M.V., Lykosov V.N. Information-Computational Technologies foe Environmental Science: Preparation of Young Researchers. Computational Technologies, 2006. V.11, Special Issue 1, p. 3-15.

[3] MIKE 2.0, Big Data Definition. (http://mike2.openmethodology.org/wiki/Big_Data_Definition)

[4] Dan Kusnetzky. What is "Big Data?". ZDNet. (Электронный ресурс: http://www.zdnet.com/blog/virtualization/what-is-big-data/1708)

[5] Ashley Vance. Start-Up Goes After Big Data With Hadoop Helper. New York Times Blog. (Электронный ресурс: http://bits.blogs.nytimes.com/ 2010/04/22/start-up-goes-after-big-data-with-hadoop-helper/?dbk).

[6] Kalinichenko L.A., et al. Problems of access to data in Data Intensive Domain in Russia. Informatika i ee primeneniya, v. 10, p. 3-23, 2016.

[7] Hilbert, Martin. "Big Data for Development: A Review of Promises and Challenges. Development Policy Review". martinhilbert.net. Retrieved 2015-10-07

[8] Shekhar S. Spatial Big Data // Proc. AAG-NIH Symp. on Enabling a National Geospatial Cyberinfrastructure for Health Research. July 2012. Minneapolis. USA

[9] Gordov E.P., Lykosov V.N. Development of information-computational infrastructure for integrated study of Siberia environment. Computational Technologies, 2007. V.12, Special Issue 2, p. 19-30.

[10] Stefano Nativi, Mohan Ramamurthy, Bernd Ritschel. EGU-ESSI Position Paper. [Digital resource]. — Access : http://scert.ru/files/EGU-PositionPaper-final.pdf

[11] Steiniger S., Hunter A.J.S. Free and open source GIS software for building a spatial data infrastructure. / In: Bocher E., Neteler M., (eds.), Geospatial Free and Open SourceSoftware in the 21st Century, LNGC, Heidelberg, Springer, 2012a, p. 247-261.

[12] Koshkarev A.V., Ryakhovskii A.V., Serebryakov V.A. Infrastructure of distributed environment of storage, search and transformation of geospatial data. Open Education, 2010, No 5, p. 61-73.

[13] Krasnopeev S.M. Experience of deployment of key elements of special data infrastructure on the base of web-services. Proceedings of XIV All-Russia Conference “Internet and Modern Society”, Sankt Peterburg, Russia, 2001, p. 92-99.

[14] Koshkarev A.V. Geoportal as a tool to control spatial data and services. Spatial data, 2008, No 2, p. 6-14.

[15] Yakubailik O.E. Geoformation geoportal, Computational Technologies, 2007. V.12, Special Issue 3, p. 116-125.

[16] Dragicevic, S., Balram, S., Lewis, J. The role of Web GIS tools in the environmental modeling and decision-making process // 4th International Conference on Integrating GIS and Environmental Modeling (GIS/EM4): Problems, Prospects and Research Needs. Banff, Alberta, Canada, September 2 - 8, 2000.

[17] Frans J. M. van der Wel. Spatial data infrastructure for meteorological and climatic data // Meteorol. Appl. 12, 2005. Pp. 7-8. DOI:10.1017/S1350482704001471.

[18] Vatsavai, Ranga Raju, Thomas E. Burk, B. Tyler Wilson, Shashi Shekhar. A Web-based browsing and spatial analysis system for regional natural resource analysis and mapping // Proc. of the 8th ACM int. symp. on Advances in geographic information systems. 2000. Washington, D.C., US., p. 95-101.

[19] Katleen Janssen. The Availability of Spatial and Environmental Data in the European Union: At the Crossroads Between Public and Economic Interests / Kluwer Law International, 2010, ISBN 9041132872, 9789041132871, 617 p.

[20] J.D. Blower, A.L. Gemmell, G.H. Griffiths, K. Haines, A. Santokhee, X. Yang. A Web Map Service implementation for the visualization of multidimensional gridded environmental data // Environmental Modelling & Software, Volume 47, September 2013, p. 218–224. doi:10.1016/j.envsoft.2013.04.002

[21] L. Becirspahic and A. Karabegovic. Web portals for visualizing and searching spatial data // Information and Communication Technology, Electronics and Microelectronics (MIPRO), 2015 38th International Convention on, Opatija, 2015, pp. 305-311. doi: 10.1109/MIPRO.2015.7160284

[22] I.G. Okladnikov, E.P. Gordov, A.G. Titov, T.M. Shulgina. Information-computational System for Online Analysis of Georeferenced Climatological Data // Selected Papers of the XVII International Conference on Data Analytics and Management in Data Intensive Domains (DAMDID/RCDL 2015), October 13 – 16, 2015, Obninsk, Russia / Edited by Leonid Kalinichenko and Sergey Starkov. CEUR Workshop Proceedings. Vol. 1536. P. 76-80.

[23] Shea Frederick, Colin Ramsay, and Steve Cutter Blades. Learning Ext JS. / Packt Publishing, 2008.– 299pp.

[24] RSF Project №16-19-10257 “Virtual computational information environment for analysis, evaluation and prediction of the impacts of global climate change on the environment and climate of a selected region” (http://rscf.ru/en/enprjcard?rid=16-19-10257).

[25] Santokhee, J. Blower, K. Haines. Storing and Manipulating Gridded Data In Spatial Databases // Reading E-science Center, University of Reading

[26] Titov A.G., Okladnikov I.G., Gordov E.P. Development of web-GIS for climatic change analysis and projections on the basis of services of spatial data processing and visualization. Proceedings of XXI Baikal All-Russia Conference "Information and mathematical technologies in science and control".

[27] Ryazanova A.A., Voropay N.N., Okladnikov I.G. Application of information and computing web system «Climate» for estimationn of aridity of South Siberia. // Proc. of International Conference and Early Career Scientists School on Environmental Observations, Modeling and Information Systems ENVIROMIS-2016, July 11-16, 2016, Tomsk, Russia. Pp. 358-362.

Satellite Data Infrastructures

© Efim Kudashev © Alexander Belov
© Natalya Kalenova
Space Research Institute (IKI), Moscow
kudashev@iki.rssi.ru , belovaf@gmail.com, kalenova7@gmail.com

Abstract

In this paper, we focused on some recent advances to the research of Earth Observation (EO) data from satellites and evolution the EO Data Archives. We review existing methods for Spatial Data Infrastructure (SDI) development based on the integration of spatial information and services. A comprehensive analysis of the problem of creating SDI is beyond the scope of this paper. Next, we will concentrate on forming SDI regarding creation of Russian segment for improving the accessibility of Russian satellite data.

1 Remote Sensing Satellite Data

Remote Sensing from satellites allow a global perspective on Earth Observation to be developed. Big Data from Space is an emerging domain given the recent sharp increase of information. Fortunately, this increase is paralleled by tremendous amount of new developments related to big data in other fields and enabled by technological breakthroughs and new challenges in hardware and software developments, multi-temporal data analysis, data management and information extraction technologies. In addition, the recent multiplication of open access initiatives to big data from space is giving momentum to the field by widening substantially the spectrum of users as well as awareness among the public while offering new opportunities for scientists and value-added companies. This is especially true for Satellite data with the public release of the complete archive of Landsat data by the United States Geological Survey. At an even larger scale, the ambitious and unique European Union Copernicus Program whose Sentinel missions operated by ESA will deliver free and open access to global data in the microwave and optical/infrared ranges. The focus is on the whole data life cycle, ranging from data acquisition by space borne and ground-based sensors to data management, analysis and exploitation in the domains of Earth Observation.

We consider some recent advances to the research of Earth Observation data from satellites and evolution the EO Data Archives [1, 7]. After a long, slow rise since 1986, the volume of satellite data from ESA's satellites passed three PB and is in constant increase. New datasets, coming from the future Earth Explorer Copernicus missions will contribute to increase the volume in the years to come. Copernicus is the new name for the Global Monitoring for Environment and Security programme, previously known as GMES. Copernicus will provide information services to support policies related to environment and security at European level, but with the broader objective of operational services on a global scale. In 2015 an average of 3 TB of satellite data was generated daily. By the end of 2016 it is projected that this figure will increase to more than 6 TB a day [2]. Prediction of GSCB (Ground Segment Coordination Body) is that the volume of satellite data will exceed 20 PB for 2020.

Example 1. Russian Satellite Data Center at Vladivostok. The Pacific Center provides the studies on physics of the ocean and atmosphere and concludes reception, processing and archiving data from satellites AQUA, TERRA, MTSAT-1R, FY-1D and NOAA.

Example 2. German Aerospace Center (DLR) created the National Satellite Data Archive. The Earth Observation Center (EOC) at DLR is the Center of competence in Germany, providing expertise in Earth Observation research and development activities, as well as operational tasks for data reception, processing and archiving. The powerful and centralized archive at the DLR Earth Observation Center has proven its stability and flexibility to allow Long Term Data Preservation over more than 20 years with nearly exponentially growing data capacity. In 2012, input/output data rates have grown to be beyond 100 MB/s, but the disk drives and networks have also grown. Archiving capacity of National Satellite Data Archive which is Remote Sensing Data Center of the Federal Republic of Germany is 2.2 PB [3].

The ESA Earth Observation Ground Segment Department operates the Earth Observation Research and Service Support (RSS). RSS primary mission is to support the EO data user's community, to ease the development of applications adding value to raw data. The RSS environment also serves the ESA harmonization activities, collecting and classifying ground segment technology development needs. With the launch of the Sentinel-1, the RSS data farm will increase the flow rate of data by a magnitude of 100, to 1 TB of data by day.

Proceedings of the XVIII International Conference «Data Analytics and Management in Data Intensive Domains» (DAMDID/RCDL'2016), Ershovo, Russia, October 11 - 14, 2016

Growing satellite data in USA. The volumes of NASA and NOAA archives have grown from 1 PB in 2000 to 10 PB in 2011 annually. Total volume of NOAA Archives will exceed 100 PB [4].

2 Spatial Data Infrastructure (SDI)

A few years ago, an idea for creating Spatial Data Infrastructure (SDI) arose. SDI is to become as an aggregate of standards for interaction of open systems, technological solutions, human resources and legal agreements for collecting, processing, distributing and using spatial data. SDI is a framework of spatial data, metadata, users and tools that connected in order to use satellite data in an efficient and flexible way. SDI is an infrastructure for spatial and digital cartographic data, which joins information resources and metadata in a GIS portal. The portal provides access to the metadata, description-based search of metadata and delivery of data. GIS portal is an integrated node of access to spatial data, which is independent of data location, format and storing structure. The development of the e-Infrastructure SDI began in the last years within the framework of INSPIRE program (Infrastructure for Spatial Information in Europe) [5]. The INSPIRE develops distributed infrastructure for geographical data for the protection of the environment in Europe, monitoring of natural resources and natural catastrophes. The geoportal INSPIRE has been created in 2005 under the European Commission initiative. The components of INSPIRE are metadata, collections and data processing services; network services and technologies; agreements on distribution, access and usage of data; mechanisms for coordinating and monitoring. INSPIRE aims at making available relevant, harmonized and quality spatial data; integrating other standards during the development process facilitates interoperability. INSPIRE infrastructure addresses technical standards, OpenGIS specifications, and protocols, including data access and the creation and maintenance of spatial information.

3 Integration of Spatial Data and Services on SDI and INSPIRE programs

The problem of data exchange is one of the most important problems in the development of SDI. This is due to difficulties in getting and processing satellite data within single infrastructure. First, this problem is related to various possibilities and requirements of providers and consumers and a lack of single standard and approach to the resources delivery. Information systems, which provide transmission on the Web of the satellite data are being intensively developed on the foundation of SDI and INSPIRE programs. The integration of spatial information and services is formed with uniform standards and protocols for data exchange. The preferred standards are ISO/TC 211 19115, CEN, OGC and W3C. The common goals of SDI and INSPIRE programs are providing coordinated distributed access to satellite informational resources; supporting solution of fundamental and applied problems related to Earth Observation from Space. The European Union's INSPIRE program is ambitious international project, which can serve as an example of implementation of the SDI, including space data.

Russian Federation began creating spatial data infrastructure for electronic exchange of spatial data and distributed access to cartographic products over the Web. The concept of *e-infrastructure* for spatial data, developed by Russian Cartography [6], defines SDI *e-Infrastructure* as a system, which provides interaction for end-users using various digital spatial data in their work. The goal of creating SDI is forming a consolidated information environment, which provides search, publication and exchange of various geographical information resources. SDI is a hierarchically arranged system build on informational technologies and based on common standards for spatial data and metadata. SDI also consists of a network of geographical information nodes — geoportals and metadata catalogues. A geoportal is a core of the informational infrastructure; it is an essential element for internal and external data exchange. A portal is a key constituent of infrastructure; it is an informational node, which contains interface for access to database (issue-related collections of satellite data, informational products of space monitoring) and a metadata catalogue.

4 Service Support Environment

Service Support Environment (SSE) is Open SOA-based environment for users and providers of data and services based on EOLI-XML, using SOAP message exchanging protocol and WSDL [7, 8]. EOLI (Earth Observation Link) is the European Space Agency's client for Earth Observation Catalogue and Ordering Services. It is used as the primary interface. EOLi works to provide browsing of the metadata and preview images of Earth Observation data acquired by the satellites: Envisat, ERS, Landsat, IKONOS, DMC, ALOS, SPOT, Kompsat, Proba, IRS, and SCISAT. SSE applies a full service oriented architecture (SOA) based on the recognized Web service standards. The system was implemented in accordance with the standards from the W3C, OASIS and the Open Geospatial Consortium organizations. Services SSE is based on different technologies that have been integrated into a unique and powerful environment. This approach allows the service provider to offer the functionalities either as an individual Web Services or aggregated at a single interface. The main aims of creating SSE are:

- give service providers an environment that would at most simplify interaction such as “service-provider — user” and “service-provider — service-provider”;
- simplify integration of existing services, providing universal XML-based interface, allowing for keeping their structure;
- provide the users with a single entry point — Internet- portal.

The core of the system consists of two main components — SSE Portal Server and AOI Server, which together form an Internet-portal, with which a user interacts. SSE Portal Server provides a web interface for users to access the portal. AOI Server works together with the SSE Portal Server and is used for giving service providers the functions of visualizing search results and of visual selection of an area on the map (Area of Interest, AOI) when specifying search criteria. Since the majority of services requires specifying geographical area as one of the search parameters, SSE Portal provides special support of this feature. When forming a query, user should specify a geographical area. User can do it in several ways:

1. choose from a list,
2. specify an area on the map,
3. upload a file, describing AOI,
4. specify an area on the map AOI by providing coordinates explicitly.

SSE Portal processes user's request using applet, loaded from AOI server, for displaying selected area on the map. SSE can visualize both locally stored information and additional map layers, downloadable from remote OGC WMS servers. The description of the selected area is stored in GML (Geography Markup Language) format and is sent to the service provider as a SOAP-message.

5 Forming the Russian Segment of SDI

The structure of SSE was initially designed based on principle "one service provider — one data array". While forming the Russian segment of SDI, another approach is used based on what has become traditional three-level information system: data level, computations level, knowledge level. Using three-level scheme is a promising way of building information environments for supporting the Earth Observation. A new feature was implemented in the scheme of accumulating and exchanging data in the Russian segment of the SDI: appearance of conceptually different components — so-called —"network nodes" — which aggregate data from several service providers without the necessity of installing and setting-up additional software. In such a connection scheme all the required data conversion to a single standard is realized in one node, while in a standard scheme this work needs to be done individually by each service provider. Besides, a user will be able to search in all the data arrays of all the service providers, which connected to the system. This system uses modern technologies, standards and open-source software, which is necessary to solve all the problems on all the development levels. These are the standard for geographical metadata of ISO series 19115-2, widely acknowledged technologies WSDL, SOAP and WMS, which are the basis for the search engine, and Eclipse development environment.

Russian segment on the SDI infrastructure constructed in IKI, which currently joins two independent satellite data providers: the information on terrains of the Pacific region supplied by IAPU at Vladivostok and satellite data for terrains of Western and Eastern Siberia supplied by ICT (Institute of Computational Technologies) at Novosibirsk. The realized solutions are well scalable and they allow for joining a large numbers of service providers of heterogeneous data. To provide users' access to satellite data, ICT creates Siberian node for collecting, storing and processing satellite data. The main functions of the node are providing telecommunications for data collecting, archiving raw data, preliminary data processing, cataloging processed data, providing long-term storing of processed data, providing access to data and topical data processing a catalogue was created based on data storing system of ICT; the catalogue is regularly updated with SPOT 4 data. The catalogue also includes archive data of Russian territory of years 1982–2002 from Landsat series satellites. A service-oriented system is created based on this catalogue; the system is realized as a basic set of applications, working in the environment of the Tomcat application server. The user interfaces are developed using PHP and JavaScript technologies. Central Authentication Server module used to provide access to the system. CAS allows for creating multi-level system of user rights and access levels. The location of the receiving complex ensures receiving data, which cover Siberia, parts of Far East and Yakutia and lands of Ural and Central Russia. It is possible to receive data from other currently active platforms. The designed software of a network node can connect data providers by using various protocols and data exchange interfaces: SOAP, ODBC, and FTP.

6 Functionality of Russian segment

Let us consider the interaction scheme of two Satellite Centers with a Russian segment node in IKI, which use different methods of data transmission. The program-hardware complex of a segment node performs the following tasks:

- direct interaction with the SSE server, processing its search queries and returning the results (message exchange with the server by XML over SOAP);
- receiving, converting and storing metadata in the local database from the satellite centers which are part of the segment: data exchange with ICT — through HTTP gateway; data exchange with IAPU— via ODBC interface;
- Providing Web interface to a node for local search within the segment without using SSE portal.

The node uses its own DBMS, which is needed for storing metadata and local collections of each satellite center, which is connected to the node. Satellite data themselves are not stored on the node server; instead, only links to them are in the local bases, and direct access provided via these links, when necessary. The metadata in the node database is updated automatically with

specified periodicity; they synchronized with local databases of each satellite center. All the necessary conversion is done in the process, since every provider stores its data in a proprietary format. Apart from the local mirror of the metadata database, each of the data providers, which is connected to the server, corresponds to its own program module, which is used for receiving and processing of the requests and for forming the response with the search results.

The results of queries processing are consolidated, converted to a given format and sent back to SSE server. The user who sent the query gets the response as a list where he/she can select necessary data using preview and then get direct links to the data.

Conclusion

The value of big data from space depends on our capacity to extract information and meaning from them. The problems of data access and data processing issues, data exchange are one of the most important problems in the development of SDI. This is due to difficulties in getting and processing satellite data within single infrastructure. Improving the accessibility of satellite data could make a very substantial contribution to environmental monitoring. IKI experience of participation in European APARSEN project confirms the interest in satellite data [9, 10]. The initial stages of our research were carried out by IKI and IAPU in the INTAS IRIS Project: Integration of Russian Satellite Data Information Resources with the Global Network of Earth Observation Information Systems. Authors have been studied integration of Russian Satellite Data with the Global Network of Earth Observation Information Systems. In this paper, we describe a distributed infrastructure for satellite data. The Russian segment of distributed informational system has been built based on EOLI-XML and SSE technologies. Technologies and principles of building distributed systems considered in this article have great potential for further development, and even today, they provide extensive means for arranging distributed heterogeneous data and services in a single global system.

References

[1] Bartunov O., Kalenova N., Kudashev E. Earth Observation Data and Creation of e-Infrastructure for Scientific Data // In ESA 2014 Conference on Big Data from Space (BiDS'2014). Conference Proceedings. European Commission, Joint Research Centre. Doi: 10.2788/1823. http://esaconferencebureau.com/2014-events/BigDatafromSpace/introduction

[2] C. Reck et all. German Copernicus Data Access and Exploitation Platform. In ESA Conference on Big Data from Space BiDS'16, Spain, 2016. http://esaconferencebureau.com/custom/16M05/bids/ALL/D2_0920_Reck.pdf

[3] C. Reck et al. Behind the Sciense at the DLR National Satellite Data Archive. In PV 2011 Conference Ensuring Long-Term Preservation and Adding Value to Scientific and technical data. CNES. Toulouse, France.2011.

[4] Ramapriyan H.K. Development, Operation and Evolution of EOSDIS - NASA's major capability for managing Earth science data. Workshop on Repositories in Science & Technology: Preserving Access to the Record of Science. 2011. http://www.cendi.gov/presentations/11_30_11_Ramapriyan_NASA_EOSDIS.pdf

[5] The INSPIRE Directive: a brief description. https://inspire.jrc.ec.europa.eu

[6] Development of SDI for Russian Federation. https://rosreestr.ru/activity/infrastruktura-prostranstvennykh-dannykh

[7] ESA Earth Online. How to Access EO Data https://earth.esa.int/web

[8] I. V. Nedoluzhko, O. O. Korobkova. Means used for integration of catalogues in modern European EO infrastructures // Russian Digital Libraries Journal. 2012. Issue 3. http://www.elbib.ru

[9] APARSEN project - Alliance for Permanent Access to the Records of Science in Europe Network. http://www.alliancepermanentaccess.org/index.php/about-aparsen/

[10] David Giaretta. Alliance for Permanent Access. http://www.alliancepermanentaccess.org/index.php/community/conferences/apa-conferences/apa-conference-oct-2014/

Росгидромет как цифровое предприятие

© Е.Д. Вязилов
ФГБУ «ВНИИГМИ-МЦД»

г. Обнинск

vjaz@meteo.ru

Аннотация

Показаны перспективы развития Росгидромета, как цифрового предприятия. Рассмотрены существующие средства учета наблюдательных подразделений, сбора, первичной обработки данных, учета и доставки информационной продукции, мониторинга автоматических комплексов измерений, сбора, обработки, прогноза и доведения информации.

1 Введение

В области гидрометеорологии произведена практически 100% оцифровка основных видов данных, собираемых как в режиме реального времени, так и в отложенном режиме. В Росгидромете автоматизированы основные процессы сбора, первичной обработки, подготовки ежемесячников и ежегодников, анализа, прогноза, контроля, хранения, обмена, доступа и визуализации данных. В рамках проекта «Модернизация Росгидромет-1» (2008-2013 гг.) приобретено более 1500 современных автоматических комплексов (АК), передающих информацию без участия человека. Кроме того, в рамках Федеральной целевой программы «Геофизика» разработано и установлено более 500 современных геофизических приборов. Использование данных с датчиков в реальном режиме времени позволяет иметь данные о гидрометеорологической и геофизической обстановке с дискретностью 10 мин и менее. Появилась возможность интеграции данных, предоставляемых различными организациями Росгидромета и других ведомств [4]. Все это открывает новые возможности по развитию гидрометеорологического обеспечения промышленных предприятий и населения.

Труды XVIII Международной конференции DAMDID/RCDL'2016 «Аналитика и управление данными в областях с интенсивным использованием данных», Ершово, 11-14 октября 2016

Несмотря на успехи Росгидромета в области применения информационных технологий для сбора, обработки, хранения, межведомственного и международного обмена данными, автоматизация управления гидрометеорологической службой сводится пока к использованию отдельных автономных систем, например, электронный документооборот, бухгалтерские системы, различные правовые, нормативно-методические и другие системы. В тоже время имеется насущная необходимость автоматизации управления процессами обработки данных, охватывающими все этапы жизненного цикла данных - от наблюдения до их использования, осуществляемые в едином информационном пространстве.

В статье делается попытка представить развитие Росгидромета как цифрового предприятия [7]. Термин «цифровое предприятие» пока не устоялся. В статье используется следующий термин [8]. Цифровое предприятие — это организация, объединяющая географически разделенные субъекты, которые взаимодействуют в процессе производства наблюдений, широко используя преимущественно электронные средства коммуникаций и информационные технологии (ИТ) во всех сферах своей деятельности, начиная от сбора данных, и заканчивая доведением информационной продукции (ИП), которая тоже представляется в цифровом виде.

Целью развития Росгидромета как цифрового предприятия является существенное повышение уровня автоматизации, начиная от сбора данных и заканчивая их использованием на промышленных предприятиях. Основными задачами управления Росгидрометом как цифровым предприятием должны стать:

- планирование работ Росгидромета и развитие наблюдательных сетей;
- оптимизация состава наблюдательной сети с использованием экономических (стоимость содержания сети) и методических (плотность наблюдательных пунктов) критериев;
- оценка финансовых затрат на поддержку наблюдательных подразделений - НП (фонд зарплаты, транспорт, связь) по

территориальным управлениям Росгидромета, типам НП;

- уменьшение дублирования при подготовке ИП и повышение уровня автоматизации при ее создании;
- автоматизация доставки ИП, необходимой пользователям и включающей не только цифровые данные в виде текста, карты, графика, таблицы, но и сведения о воздействиях гидрометеорологических условий на промышленные объекты и население, рекомендации для принятия решений;
- оптимизация решений, принимаемых пользователями на основе гидрометеорологической информации, оценка возможных убытков и расчет стоимости превентивных мероприятий;
- обеспечение надежности работы аппаратно-программных комплексов (АПК) сбора, обработки, прогноза и доведения до пользователей гидрометеорологической информации, и выявления проблем при эксплуатации элементов инфраструктуры доведения информации;
- подготовка ежегодных отчетов, в состав которых должны входить показатели состояния наблюдательных платформ (количество научно-исследовательских судов, их средний возраст, суммарное водоизмещение, количество прибрежных станций по годам, средняя плотность станций на 100 км побережья, др.).

Можно привести пример реализации некоторых идей цифрового предприятия в области наук о Земле. Так Геофизический центр РАН и Институт программных систем РАН создают Виртуальную обсерваторию для метаресурсов научных данных на основе GRID технологии [2].

Большие работы ведутся в этом направлении за рубежом. Отмечается [1], что до 8% предприятий в США и 5% Западной Европы уже работают как цифровые предприятия. В Национальной администрации США по океану и атмосфере ведутся большие работы по автоматизации большинства перечисленных функций, но она пока не ставят вопрос о создании цифрового предприятия. Там видят развитие гидрометеорологической службы в создании систем «Виртуальная Земля», «Виртуальный океан».

Автор статьи не претендует на новую концепцию цифрового предприятия, здесь показывается, как Росгидромет двигается к цифровому предприятию. В статье рассматривается общее видение цифрового предприятия и даны краткие сведения о средствах учета производства наблюдений, выпускаемой информационной продукции, мониторинга состояния аппаратно-программных средств сбора и обработки данных.

2 Общее видение цифрового предприятия

Современное традиционное предприятие, как правило, имеет вычислительную сетевую среду, созданную в рамках одной организации для выполнения основных ее функций, с внешней связью через Интернет. В цифровом предприятии приложения и ресурсы выходят за рамки отдельной организации [7]. Конечно, чтобы создать цифровое предприятие, требуются средства управления вычислительными, информационными, сетевыми ресурсами и создание единой системы информационной безопасности.

Приложения, ориентированные на Интернет, беспроводные среды, аренде программных средств, доступ к распределенным информационным ресурсам (ИР) и другие достижения ИТ–технологий позволяют создать цифровое предприятие уже на имеющихся программно-технических средствах. Резкое увеличение пропускной способности сетей позволяет аккумулировать данные на централизованных серверах цифрового предприятия и обеспечивать удаленный доступ к данным для любого пользователя.

Понятие «цифровое предприятие» означает информационную прозрачность, что достигается наличием систем учета на всех этапах жизненного цикла обработки данных от измерения на наблюдательном пункте до использования информации при поддержке решений, а также наличием бизнес-процессов управления Росгидрометом, включая функции управления доставкой и использованием информации на промышленных предприятиях.

Основной целью работы Росгидромета как цифрового предприятия является повышение эффективности использования гидрометеорологической информации руководителями органов государственной власти, промышленных предприятий и населением.

Решение всего комплекса задач контроля и управления АПК как цифрового предприятия должно обеспечиваться системой управления, которая предназначена для оперативного обеспечения Центрального аппарата (ЦА) и организаций Росгидромета данными и информацией о состоянии НП, сбора данных, ресурсов, АПК, ИП.

Любое управление предприятием основано на системах учета ресурсов: финансовых, кадровых, производственных (помещения, электроэнергия, транспорт, связь, пользователи), наблюдательных (пункты наблюдений), телекоммуникационных, вычислительных, информационных.

В Росгидромете еще в конце восьмидесятых годов прошлого столетия была начата работа по созданию автоматизированной системы учета наблюдательных подразделений (АСУНП, http://cliware.meteo.ru/goskom_cat/list/index.jsp) [6]. Одной из решаемых задач этой системы является отслеживание состояния наблюдательных сетей.

Во многих организациях гидрометеорологической службы уже практикуется учет и мониторинг состояния телекоммуникационных, вычислительных и информационных ресурсов. При этом реализуется мониторинг ресурсов как в одной, так во многих организациях, например, в межведомственной Единой государственной системе информации об обстановке в Мировом океане (ЕСИМО) [4]. Необходимо реализовать ведомственный мониторинг таких ресурсов, чтобы ЦА Росгидромета мог видеть состояние ПН, сбора, обработки и доведения информации до пользователей.

Финансовые и кадровые ресурсы на этом фоне находятся на более низком уровне автоматизации. То есть в каждой организации Росгидромета используется своя бухгалтерская система на основе технологической платформы 1С, с помощью которой готовятся регулярные годовые и квартальные отчеты, которые отправляются в вышестоящие органы управления. На основе такой отчетной информации нельзя получить детальную финансовую информацию по основной структурной единице Росгидромета – ПН.

С помощью Интегрированной информационно-технологической системы (ИИТС), разрабатываемой в Росгидромете, и ЕСИМО [4] руководители предприятий получают данные из внешних источников.

С появлением смартфонов и планшетов цифровое предприятие означает и другой подход к гидрометеорологическому обеспечению. В нынешних условиях руководители промышленных предприятий должны автоматически получать информацию о резких изменениях гидрометеорологической обстановки. Это невозможно без автоматизации доставки сведений об опасных явлениях (ОЯ), удаленного доступа к основным параметрам гидрометеорологической обстановки и к аналитической информации.

К цифровому предприятию ведет также тотальная автоматизация производства наблюдений, включая работу автоматических комплексов измерений, и технологических процессов обработки данных. Сведения о средствах производства наблюдений, изменения в составе и условиях наблюдений должны попадать в АСУНП по месту и времени своего возникновения. Мониторинг состояния сбора и первичной обработки гидрометеорологических данных также должен быть организован с помощью АСУНП.

Важными составными частями управления цифровым предприятием являются средства обеспечения своевременности, полноты сбора и первичной обработки данных; выпускаемой ИП (бюллетени, анализы, прогнозы, обобщения); учета пользователей и обеспечения их ИП; мониторинга работы АПК, выпуска прогнозов и доведения до пользователей гидрометеорологической информации.

Использование различных систем учета, мониторинга состояния всех инфраструктурных элементов Росгидромета следует рассматривать только как первый этап построения цифрового предприятия. Огромные объемы поступающих данных необходимо учитывать в режиме реального времени, обрабатывать, сопоставлять с данными из учетных систем, систем планирования и других приложений. Фактически речь идет не только об интеграции гидрометеорологических данных, но и других видов данных (кадровых, финансовых, технических, телекоммуникационных). Такая интеграция данных позволит ЦА Росгидромета видеть аналитику о деятельности Росгидромета и принимать решения на ее основе в режиме реального времени.

Цифровое предприятие Росгидромета должно быть создано в одном из имеющихся центров обработки данных. Концепция цифрового предприятия предполагает представление некоторого набора средств (систем хранения, вычислительных мощностей, сетевых служб) в виде легкодоступных ресурсов. Внутри цифрового предприятия приложения должны строиться на базе компонентов и сервисов, которые отвечают потребностям пользователей, взаимодействующих с множеством ИР.

Первый этап создания цифрового предприятия сводится к созданию единого реестра ИР и ИП. На следующем этапе должна появиться возможность доступа к интегрированным ИР и ИП. На третьем — пользователь будет автоматически получать необходимую информацию в соответствии с заранее определенными им требованиями. Примерная функциональная схема цифрового предприятия и содержание разделов даны на рис.1.

Интегрированные ИТ-ресурсы можно искать через единый реестр ИР. Системы учета должны взаимодействовать как между собой, так и с аналитическими системами. Поэтому во всех системах учета и аналитических системах должна использоваться единая НСИ, организованная в виде БД и включающая единый словарь параметров, общие коды и классификаторы, различные словари и справочники, метаданные.

Основными преимуществами цифрового предприятия являются:

- **масштабируемость** - добавление новых серверов БД и серверов приложений позволяет поддерживать большое число параллельных соединений с клиентами, распределять нагрузки среди множества физических серверов, виртуальных машин;
- **возможность модификации бизнес-логики** для управления цифровым предприятием - создаваемая архитектура позволяет изменять бизнес-логику и модифицировать ее, не затрагивая клиентские системы;
- **многократное использование сервисов -** однажды созданные сервисы могут быть использованы другими приложениями, что

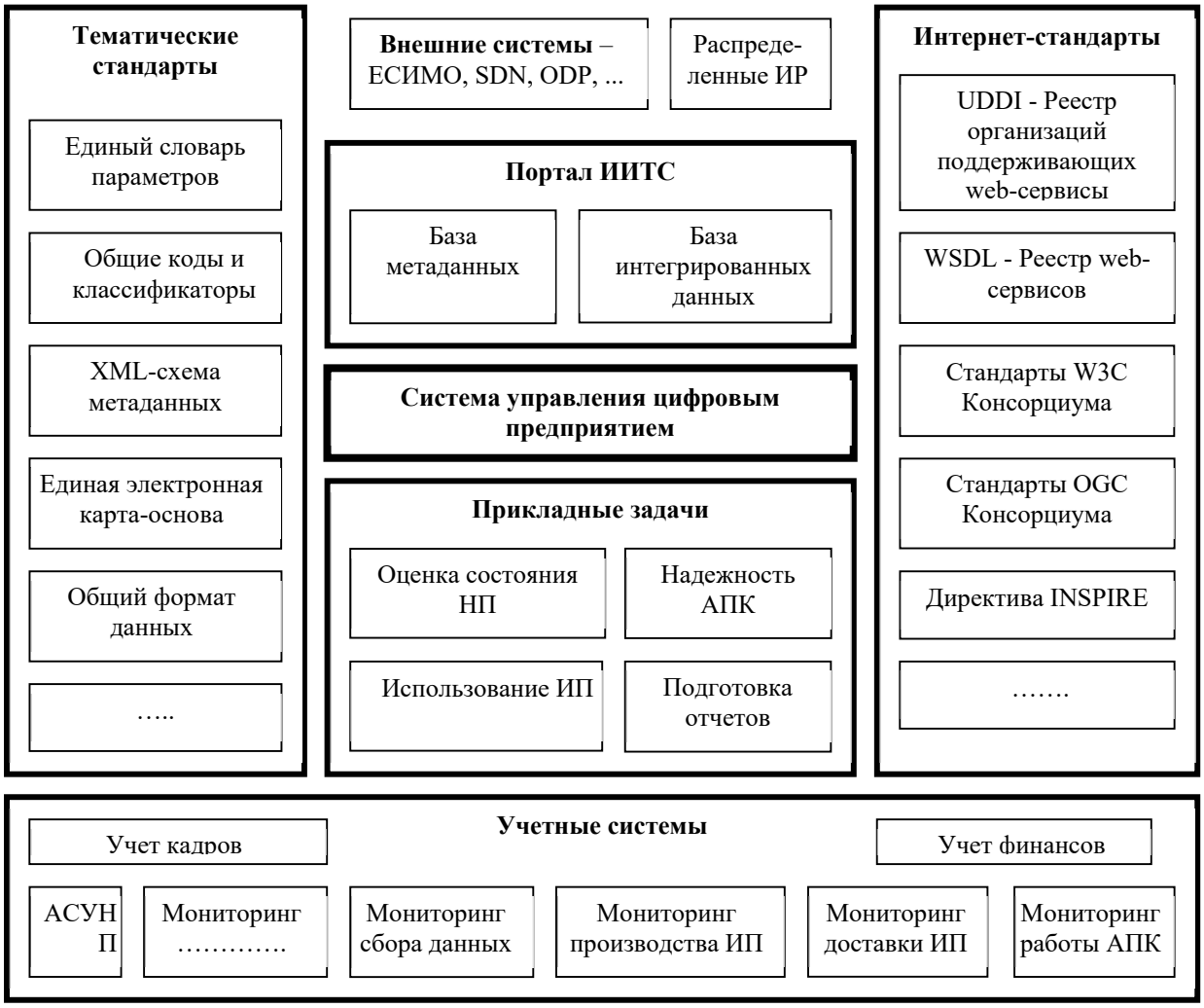


Рисунок 1 Схема функционирования Цифрового предприятия

снижает затраты на разработку и администрирование системы;

- **сохранение инвестиций** – более эффективное использование имеющихся аппаратно-программных средств.

Основные функции Росгидромета (сбор, занесение на носители, каталогизация, хранение, прикладная обработка, обмен, использование данных) должны переместиться в распределенную среду. Управление цифровым предприятием должно находиться в одном месте, а выполнение функций - децентрализовано. Кроме того, возникает возможность автоматического обслуживания пользователей по каждому населенному пункту на основе аналитических и прогностических полей высокого разрешения (до 3-5 км). Переход на автоматическое обслуживание гидрометеорологической информацией в электронном виде — это серьезная трансформация существующих технологий прикладной обработки, диагноза, прогнозирования, доведения и визуализации ИП, подготавливаемой Росгидрометом, и использования информации при поддержке решений на промышленных предприятиях. Перспективы такой трансформации напрямую зависят от готовности руководителей промышленных предприятий к восприятию такой системы обслуживания, а также руководителей организаций Росгидромета, включая его ЦА, к восприятию новой парадигмы гидрометеорологического обслуживания предприятий и населения – переход на персонализированное, автоматическое доведение информации на любое мобильное интернет-устройство.

Реализация цифрового предприятия создает программную среду, способную интегрировать информационные измерительные системы, системы связи, контроля гидрометеорологической обстановки; использовать структуры, протоколы и алгоритмы, присущие информационным вычислительным сетям; различные средства связи. Цифровое предприятие должно упростить обмен цифровой информацией. Такая система основывается на открытых стандартах.

Рассмотрим направления развития средств учета как составных частей системы управления цифровым предприятием.

3 Учет средств производства наблюдений

В настоящее время в России действует более 6500 гидрометеорологических станций, в т. ч. более 1900 метеорологических, 3500 гидрологических станций и постов. Этими основными объектами поставки информации нужно управлять с точки зрения уменьшения затрат на производство наблюдений, техническое оснащение.

Существующая схема управления опирается на отдельные «островки» автоматизации в виде учетных файловых систем в Excel или Word. Несмотря на то, что еще в 1988 г. была заложена основа АСУНП, по-настоящему она не использовалась для управления в силу ряда причин. Недостаточное качество созданной БД приводило к тому, что получаемая из АСУНП информация расходилась с информацией, представляемой из других источников.

В 2015-2016 гг. произведено развитие АСУНП (http://asunp-prototype.meteo.ru/portal/asunp/). Кроме устранения недостатков системы, существенно улучшен интерфейс поиска НП, реализована система удаленного ввода сведений о НП, разработаны средства агрегации данных о состоянии, обеспечении и модернизации НП по

4 Учет средств сбора и первичной обработки данных

Эти средства частично присутствуют в соответствующих системах сбора данных. Наиболее автоматизированным вариантом является система учета и мониторинга регистрации оперативных данных, поступающих по каналам глобальной сети телесвязи (ГСТ). Здесь имеется два уровня контроля поступления данных. Первый - реализован на уровне автоматизированной системы передачи данных (АСПД). На гидрометеорологические фиксированные станции, не передавшие свои наблюдения в срок в соответствии с регламентом, посылается запрос. С помощью систем ОМЕГА и Cliware (http://cliware.meteo.ru/meteo/), разработанных во ВНИИГМИ-МЦД, можно получить карту НП, передавших данных за любой период времени. Слабым местом такой схемы учета поступлений оперативных данных является отсутствие возможности контроля поступлений от подвижных наблюдательных платформ. Этот недостаток частично реализуется на основе Глобального центра информационной системы ВМО (ГЦИС), где на основе обмена с другими ГЦИС можно получить недостающие сообщения (http://portal.gisc-msk.wis.mecom.ru/portal).

Учет поступивших комплектов данных, поступающих в отложенном режиме с гидрометеорологических, гидрологических, прибрежных, агрометеорологических и других типов станций ежемесячно отмечается в текстовом редакторе. В случае, если данные не поступили вовремя, то пишется письмо в УГМС о необходимости ускорить поступление отсутствующего комплекта данных. Недостатком такой системы учета является отсутствие возможности автоматического уведомления о поступлении данных и запроса на получение отсутствующих комплектов.

Сведения о выполненных морских экспедициях отмечаются в Каталоге рейсов научно-исследовательских судов, который функционирует уже более 30 лет и насчитывает более 34000 экспедиций (http://portal.esimo.ru/portal/portal/esimo-user/metadata). Учет поступления данных может быть реализован на основе анализа заявок на проведение научных исследований, формируемого на их основе Плана экспедиционных работ на конкретный год и каталог выполненных рейсов.

В системах учета поступления данных не хватает автоматизации контроля поступления данных, средств учета передачи данных в Государственный фонд по гидрометеорологии на длительное хранение.

5 Учет выпускаемой информационной продукции

Системы мониторинга должны включать не только классические системы управления ресурсами цифрового предприятия, но и системы учета ИП, пользователей информации. Кроме списков ИП (бюллетеней, анализов, прогнозов, обобщений, справочников, атласов), периодически подготавливаемых Росгидрометом, не существует системы учета ИП, подготавливаемой во всех организациях Росгидромета. На сайте Росгидромета имеется раздел «Основные информационные ресурсы и продукция Росгидромета» (http://www.meteorf.ru/product/info/), в котором представлена 41 ссылка на ИП, представленную на сайтах НИУ и УГМС. Для этих ссылок нет метаданных, которые позволили бы организовать их поиск. Они отражают только очень небольшую часть ИП, которая готовится в Росгидромете.

В рамках создания ИИТС предполагается описать ИП Росгидромета. Учет ИП позволит контролировать ее актуальность, а главное - уменьшить дублирование в подготовке похожей продукции. Автор участвует в разработке ИИТС.

6 Учет доставки информационной продукции до лиц, принимающих решения

В настоящее время ИП доставляется нарочным (существует список организаций и лиц, которым доставляется ИП), по телефону, факсу (имеется бумажный журнал передачи с указанием даты и времени), электронной почте (имеются списки рассылки).

Чтобы получить значимый эффект от результатов использования гидрометеорологической информации в органах государственной власти и на промышленных предприятиях, уже недостаточно просто передать данные руководителю организации, необходимо помочь ему принять решение в той или иной ситуации. Для этого руководитель предприятия должен получить не только значения показателей гидрометеорологической обстановки в электронном виде, но и значение уровня опасности, сведения о возможных воздействиях (последствиях) опасных явлениях (ОЯ) на рассматриваемый объект, оценку возможного ущерба, рекомендации для принятия превентивных мер и их стоимость, а также средства выбора наиболее эффективных вариантов решений. То есть, необходима персонализация подхода к гидрометеорологическому обеспечению руководителей предприятий - каждому объекту свой состав показателей и свои критические значения индикаторов обстановки; автоматическая передача данных не только в центры сбора, но и потенциальным пользователям при превышении критических значений параметров.

В качестве единого корпоративного хранилища данных должна использоваться ИИТС, разрабатываемая и создаваемая на основе апробированных подходов, методов, ресурсов и сервисов ЕСИМО [4]. Имеющиеся функциональные возможности у этих систем позволяют доставить руководителям предприятий нужную информацию на любой объект, по любому району, в любой момент, в режиме реального времени получения данных.

Применение концепции цифрового предприятия поможет осуществлять мониторинг окружающей среды путем:

- настройки пользователей на объекты обслуживания, технологические процессы, уровни принятия решений, критерии ОЯ;
- аутентификации мобильного пользователя;
- определения сложившихся ОЯ для передачи сведений о них на мобильное устройство;
- доступа к данным на сервере с мобильного устройства.

Пользователь информации должен быть освобожден от рутинной работы по постоянному отслеживанию состояния гидрометеорологической обстановки, а, следовательно, у него появится больше времени для ее творческого анализа, он сможет принимать решения, используя для этого больше имеющихся сведений. За пользователем останется право производить содержательный анализ полученных вариантов решений, корректировать их, утверждать наилучший в качестве решения и контролировать их исполнение.

Пользователи сами настраивают критические значения параметров, при которых система должна их оповещать. Кроме сведений об ОЯ, доставляемых по SMS на мобильное Интернет-устройство должны доставляться сведения о возможных воздействиях ОЯ на промышленные предприятия, население и рекомендации для принятия решений. За счет такого решения существенно ускорится доставка информации и увеличится информативность руководителей промышленных предприятий.

Для оценки качества доставки должны использоваться следующие показатели: количество зарегистрированных пользователей; время доставки информации.

Автор статьи развивает новую парадигму гидрометеорологического обеспечения, полностью основанную на цифровых технологиях – от сбора до принятия решений на промышленных предприятиях.

7 Учет работы аппаратно-программных средств

Учет состояния АПК разделяют на две составляющих: мониторинг ИТ-инфраструктуры и мониторинг качества предоставления сервисов. Инфраструктурными компонентами являются оборудование (серверы, сеть и др.) и общее программное обеспечение. Целью мониторинга является оценка качества функционирования сервисов и своевременность их предоставления. Задачей мониторинга является оценка и контроль работоспособности оборудования и программного обеспечения. Метриками для мониторинга ИТ-инфраструктуры и ИТ-сервисов являются: объем потребляемых вычислительных ресурсов, занимаемой оперативной и дисковой памяти, нагрузка на сетевое оборудование и каналы связи. Метрики мониторинга используются по-разному. Провайдеру ИТ-услуг важно знать, какое количество ИТ-ресурсов необходимо иметь под рукой для предоставления сервисов заявленного уровня. На основании накопленной в системе мониторинга исторической информации должен выполняться проактивный мониторинг, позволяющий прогнозировать события.

Несмотря на внедрение систем мониторинга АПК, обеспечить круглосуточную доступность данных и приложений с коэффициентом готовности 0.95 и выше не очень просто. Как правило, на устойчивость функционирования влияет один или несколько программных компонентов. Например, наиболее проблемными комплексами в ЕСИМО при наличии объемных ИР (до 4.5 Гбайт) является ГИС (время создания глобальных слоев превышает 3 часа, время отклика больше 15 с). Решение этой проблемы обеспечивается выделением виртуальных машин для наиболее значимых программных компонентов, использованием кластерных серверов.

Для мониторинга качества предоставления сервисов важны показатели их доступности и качества их функционирования. Для этого вводятся показатель —время подготовки ИП.

8 Средства управления нормативно-справочной информацией

НСИ нужна для формирования и оперативного предоставления вместо кодовых значений

параметров полных их названий. Основными атрибутами, для которых необходимо применение общих классификаторов, являются коды стран, ведомств, организаций, наблюдательных платформ (метеорологических, гидрологических, прибрежных станций, научно-исследовательских и попутных судов, буев). Для многих из этих атрибутов имеются международные, национальные [5] и ведомственные классификаторы.

Решением проблемы унификации классификаторов является создание единой для всех организаций Росгидромета централизованной системы ведения НСИ, стандартизирующей общие коды и классификаторы. Единая система НСИ представляет собой информационную систему, обеспечивающую хранение, обработку и предоставление справочников и классификаторов. Некоторые общие решения по реализации такой системы представлены в работе [3]. В состав программного обеспечения системы должны входить инструменты ввода и редактирования справочников и классификаторов, средства поиска классификаторов, процедура формирования набора классификаторов для конкретных организаций и систем.

Для реализации этих решений необходимо разработать и принять:

- состав и структуру НСИ, системы классификации и кодирования;
- регламент ведения и сопровождения НСИ;
- регламент использования организациями Росгидромета НСИ особенно при интеграции действующих информационных систем;
- регламент обеспечения доступа пользователей к НСИ и их технической поддержки.

В настоящее время имеется система поиска классификаторов и кодов (http://portal.esimo.net/ portal/portal/tech/), отражающая только классификаторы, относящиеся к морской среде и деятельности.

9 Заключение

Для организации цифрового предприятия необходимо провести интеграцию финансовых, кадровых, информационных и других ресурсов; электронную паспортизацию объектов с указанием критических значений параметров ОЯ; мониторинг инфраструктуры, АПК; производства информационной продукции.

Переход к цифровому предприятию будет эволюционным. Стандартизация и унификация существующих систем классификации, кодирования для всех объектов Росгидромета, создание единого словаря параметров в области гидрометеорологии является частью работ по созданию единого информационного пространства.

Литература

[1] М. Баранов. Цифровое предприятие: пришло время перемен // PC Week №10 (909) 7 июня 2016. http://www.pcweek.ru/idea/article/ detail.php? ID=185915.

[2] Виртуальная обсерватория VxOware: метаресурс научных данных. Институт программных систем РАН. http://skif.pereslavl.ru/psi-info/rcms/rcms-posters.rus/virtual-observatory-plakat.pdf.

[3] Гулько Д. Е. Система нормативно-справочной информации: типовые ошибки и заблуждения // Журнал «Газовая промышленность». 2004. №6. НЦИТ «ИНТЕРТЕХ». http://www.intertech.ru/ About/ comppress.asp?filename=comppress_44.

[4] Михайлов Н.Н., Вязилов Е.Д., Воронцов А.А., Белов С.В. Единая государственная система информации об обстановке в Мировом океане и ее применение для информационной поддержки морской деятельности Российской Федерации. Труды ФГБУ «ВНИИГМИ-МЦД». - 2014. – Вып.177. с.95-118.

[5] Общероссийские классификаторы. 2016. http://www.gks.ru/metod/ classifiers.html.

[6] РД 52.04.107-86. Наставление гидрометеорологическим станциям и постам. Вып.1. – Л.: Гидрометеоиздат. 1987. – 183 с.

[7] Смирнов Н. Цифровая трансформация // Журнал «Директор информационной службы». 2014. № 12. http://www.osp.ru/cio/2014/ 12/13044350/.

[8] Уорнер М. Виртуальные организации: новые формы ведения бизнеса в XXI веке. Витцель. – М.: Добрая книга. 2005. 295 с.

Rushyd.romet as electronic enterprise

Evgenii Viazilov

The prospects of development of Rushydromet as the digital enterprise are showed. The existing tools for the integration of observational units, gathering and initial processing, accounting, and delivery of information products, automatic monitoring systems of measurement, collection, processing, forecasting and reporting of information are considered.

Data Management of the Environmental Monitoring Network: UNECE ICP Vegetation Case

G. Ososkov[1], M. Frontasyeva[2], A. Uzhinskiy[1], N. Kutovskiy[1], B. Rumyantsev[1,2,]
A. Nechaevsky[1], S. Mitsyn[1], K. Vergel[2]
[1]Laboratory of Information Technologies and [2]Frank Laboratory of Neutron Physics
Joint Institute for Nuclear Research, Dubna, Moscow Region, Russia
ososkov@jinr.ru marina@nf.jinr.ru

Abstract

A new data management cloud platform is presented. The platform is to be applied for global air pollution monitoring purposes to assess the pathway of pollutants in the atmosphere. For this purpose a set of inter-connected services and tools will be developed and hosted in the JINR cloud.

1 Introduction

Air pollution has a significant negative impact on the various components of ecosystems, human health, and ultimately cause significant economic damage. That is why air pollution is a main concern of the Doctrines of the environmental safety all over the world. Increased ratification of the Protocols of the Convention on Long-range Transboundary Air Pollution (LRTAP) is identified as a high priority in the new long-term strategy of the Convention. Full implementation of air pollution abatement policies is particularly desirable for countries of Eastern Europe, the Caucasus and Central Asia (EECCA) and South-Eastern Europe (SEE). Atmospheric deposition study of heavy metals, nitrogen, persistent organic compounds (POPs) and radionuclides is based on the analysis of naturally growing mosses through moss surveys carried out every 5 years [1]. Due to intense activity of the Joint Institute for Nuclear Research (JINR), as a coordinator of the moss surveys since 2014, Azerbaijan, Belarus, Georgia, Kazakhstan, Moldova, Turkey and Ukraine participated in the moss survey for 2015/2016. Nowadays the UNECE ICP Vegetation programme [2] is realized in 36 countries of Europe and Asia. Mosses are collected at thousands of sites across Europe and their heavy metal (since 1990), nitrogen (since 2005), POPs (pilot study in 2010) and radionuclides (since 2015) concentrations are determined. The goal of this study program is to identify the main polluted areas, produce regional maps and further develop the understanding of long-range transboundary pollution [3].

Proceedings of the XVIII International Conference «Data Analytics and Management in Data Intensive Domains» (DAMDID/RCDL'2016), Ershovo, Russia, October 11 - 14, 2016

2 Experiment and data interpretation

Sampling is carried out in compliance with the internationally accepted guidelines [4]. Such analytical techniques as AAS, AFS, CVAAS, CVAFS, ETAAS, FAAS, GFAAS, ICP-ES, ICP-MS, as well an INAA are used for elemental determination. A total of 13 elements are reported to the Atlas (As, Cd, Cr, Cu, Fe, Hg, Ni, Pb, V, Zn, Al, Sb, and N). Nowadays POPs (whichever determined) and radionuclides (namely, 210Pb and 137Cs) are accepted for air pollution characterization. The results are reported as number of sampling sites, minimum, maximum and median concentrations in mg/kg. The data interpretation is based on Multivariate statistical analysis (factor analysis), description of sampling sites (MossMet information package) and distribution maps for each element produced using ArcMAP, part of ArcGIS, an integrated geographical information system (GIS) [5]. Examples of GIS maps are presented in Fig. 1.

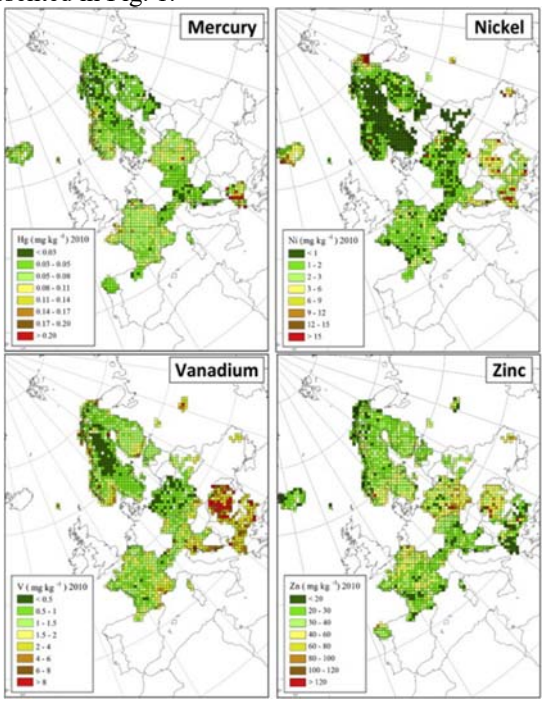


Figure 1 Examples of distribution maps [3]

Analytical results and information on the sampling sites (MossMet set) reported to JINR include confidential

acceptance of the data from individual contributors, the storage of large data arrays, their initial multivariate statistical possessing followed by applying GIS technology, and the use of artificial neural networks for predicting concentrations of chemical elements in various environments.

As an example of the importance of this study, the tendency of average median metal concentrations in moss (± one standard deviation) since 1990 to 2010 are presented in Fig. 2.

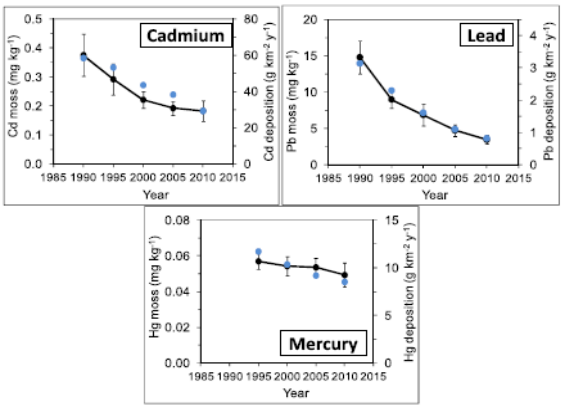


Figure 2 Change in atmospheric deposition of elements in time. The black dots in the graphs show the decline in deposition across Europe and blue dots as modeled by the Environmental Monitoring and Assessment Program

3 Motivation and aim

As discussed above, the ICP Vegetation programme is very important project, but it has a serious weakness related to its weak adoption of modern informational technologies. There are dozens of respondents in existing monitoring network and their number is increasing, but information on collecting and processing of samples is carried out manually or with minimum automation. Data mostly stored in xls files and aggregated manually by the coordinator. Files from respondents are usually passed to the coordinator by email or by ordinary mail. There are no common standards in data transfer, storing and processing software. Such situation does not meet the modern standards for quality, effectiveness and speed of research. Lack of a single web-platform that provides comprehensive solution of biological monitoring and forecasting tasks is a major problem for research.

Therefore the aim of the project is to create a cloud platform using modern analytical, statistical, programmatic and organizational methods to provide the scientific community with unified system of gathering, storing, analyzing, processing, sharing and collective usage of biological monitoring data.

The platform elements are to facilitate IT-aspects of all biological monitoring stages starting from a choice of collection places and parameters of samples description and finishing with generation of pollution maps of a particular area or state-of-environment forecast in the long term. Mechanisms and tools for association of participants of heterogeneous networks of biological monitoring are to be provided in the platform. That enables verifying obtained results and optimizes research. The open part of the platform can be used for informing public authorities, local governments, legal entities and individuals about state-of-environment changes.

One more important aspect of ecological researches relates to various statistical methods applied to process collected data. Modern approaches to explore air pollutions provided by heavy metals, nitrogen, POPs and radionuclides include as a mandatory part multivariable statistical and intellectual data processing. Latest tendencies in data processing include extension of a set of georeferenced data that is integrated in data processing of surveyed data. So it is not limited by geographical, topographical or geological information, what is traditional in such cases, but also includes, for example, satellite imagery and their products, topographic high-precision data derived from aerial photography, etc. These new data classes, contrary to the traditional ones, are characterized by a high resolution and dynamic nature – for example, satellite images represents a reflection of solar radiation, which depends on the time of day, season, cloud cover, etc. This in turn greatly increases the amount of data to be processed. The task of integration of different types of data is tied to the problem of the development of new models and algorithms – such as neural networks [6], self-organizing maps [7], etc. – during the study of dynamic properties of ecological processes among other things.

So, one more aim of our project is to develop modern software tools for multivariable statistical and intellectual data processing oriented on the GIS-technology.

4 ICP vegetation data and required resources

The moss data are to be collected for about 50 countries in Europe, Asia and Central America. Each country has more than 100 monitored points, and several hundred parameters must be taken into account for each of them. A bulky archive is needed to perform comparative studies and to estimate dynamics of explored air pollution processes. Keeping in mind the intensive data exchange and non-relational and poor-structured character of data we can assess the size of our database on the level of terabytes.

Thus it is necessary for scientists to manage large amounts of data, and it leads to many non-trivial problems in IT field. It seems natural that a solution should be centralized and outsourced to a cloud.

From a cloud point of view, the amount of data and computing leads to data intensive processing.

5 Data management on the unified cloud platform

To optimize the whole procedure of data management, it is proposed to build a unified platform consisting of a set of interconnected services and tools to be developed, deployed and hosted in the JINR cloud [6].

The JINR cloud currently has 400 CPU cores, 1000 TB of RAM and about 30 TB of total local disk space on cloud worker nodes for virtual machines and containers deployment. Hosting services in the cloud allows scaling up and down cloud resources assign to the services depending on theirs load. When some component will require more resources cloud can provide it without affecting other components. This increases the efficiency of hardware utilization as well as the reliability and availability of the service itself for the end-users. Such auto-scaling behavior will be achieved by using the OneFlow component of OpenNebula platform [6], which the JINR cloud is based on.

We define requirements for the platform and specify its components. The general architecture of the platform and technologies used are depicted in Fig. 3.

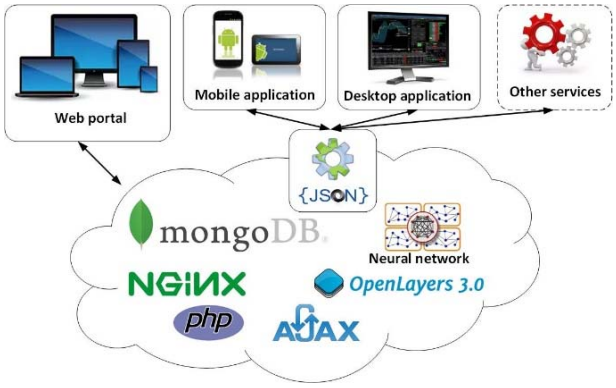


Figure 3 General architecture of the platform and technologies used

We analyze data that comes from the contributors. The data samples can have 10 to 40 metrics depending on the collecting area. Most of the metrics are optional, so traditional relation databases will be ineffective. We also want to have a possibility to change structure of the data sample object without hard code modification to easily integrate new projects and experiments into the platform. We have a positive experience with MongoDB (open-source, document database designed for ease of development and scaling [9]) at our previous projects where more than 5 million data records from 200+ contributors are processed so it was decided to use the data base to store sampling results.

The portal back-end will be built on Nginx (an open source reverse proxy web server for HTTP protocol [10]) and developed with PHP (widely-used open source general-purpose scripting language that is especially suited for web development). That should provide necessary performance and scalability. Web-portal with responsive design that adjusts to different screen sizes is the main interface of the platform. The portal allows multilevel access to the data and has advanced data processing and reporting mechanisms. Currently basic functionality of the portal has been implemented and authorized users can manage their project/regions, import data samples and generate regional maps. At top of Fig. 4 one can see the interface for project management where contributor can add, delete, edit or copy the datasets. At bottom of the figure the map with the indication of pollution distributions and basic instruments to configure the map are presented.

We have tried QGIS (Open Source Geographic Information System [11]) and OpenLayers (opensource javascript library to load, display and render maps from multiple sources) for regional and global maps representation. But QGIS and its web plugin is too hard to maintain and develop. Now we are using OpenLayers [12] and some of its specific layers that allows to do basic interpolation to create concentration maps.

Another interface to the platform is RESTful service [13] that we are going to provide to the mobile and desktop application and also for third-party services that can be interested in the environmental monitoring data.

Data import and export mechanism will be available for the platform, so users can process data online or upload it and use their local processing application. Intelligent multi-level statistical data processing is one of the platform important parts. We have tried several solutions but statistical and analytical packages are still under discussion. A very promising direction is the use of artificial neural network applications for predicting concentrations of chemical elements in various environments. We have done some research in this field but do not yet have the finished solution.

6 Prediction and GIS-oriented data processing

Prediction is an important step of data analysis of any ecological survey. Application of prediction methods enables mapping of estimate values. Maps in their turn provide visualization of spatial variability of data and can be used for visual analysis so that ecological hazards can be identified [14].

Kriging is a widely-used interpolation technique used for prediction, e.g. concentration of heavy elements in moss [15], soil contamination [6]. Recently more and more research is made towards integration of different data sources like aerial and satellite photography together with incorporation of new methods like artificial neural networks.

Mathematically, given a discrete function $f(x_i, y_i)$ (response variable defined by measurements over Cartesian coordinates) on an irregular grid of a set of points $V = \{(x_i, y_i)\}$, an interpolation procedure finds $\hat{f}(x, y)$ for f such that $\hat{f}(x, y)$ is prediction for f for $\forall (x, y) \in \mathbb{R}^2$. Integration of data is done in such a way that helps an interpolation procedure, like artificial neural network frameworks, to make a better predictor (interpolator). Such an approach is based on a conjecture that neural networks are capable to employ hidden non-linear correlations that exist and hidden in the data.

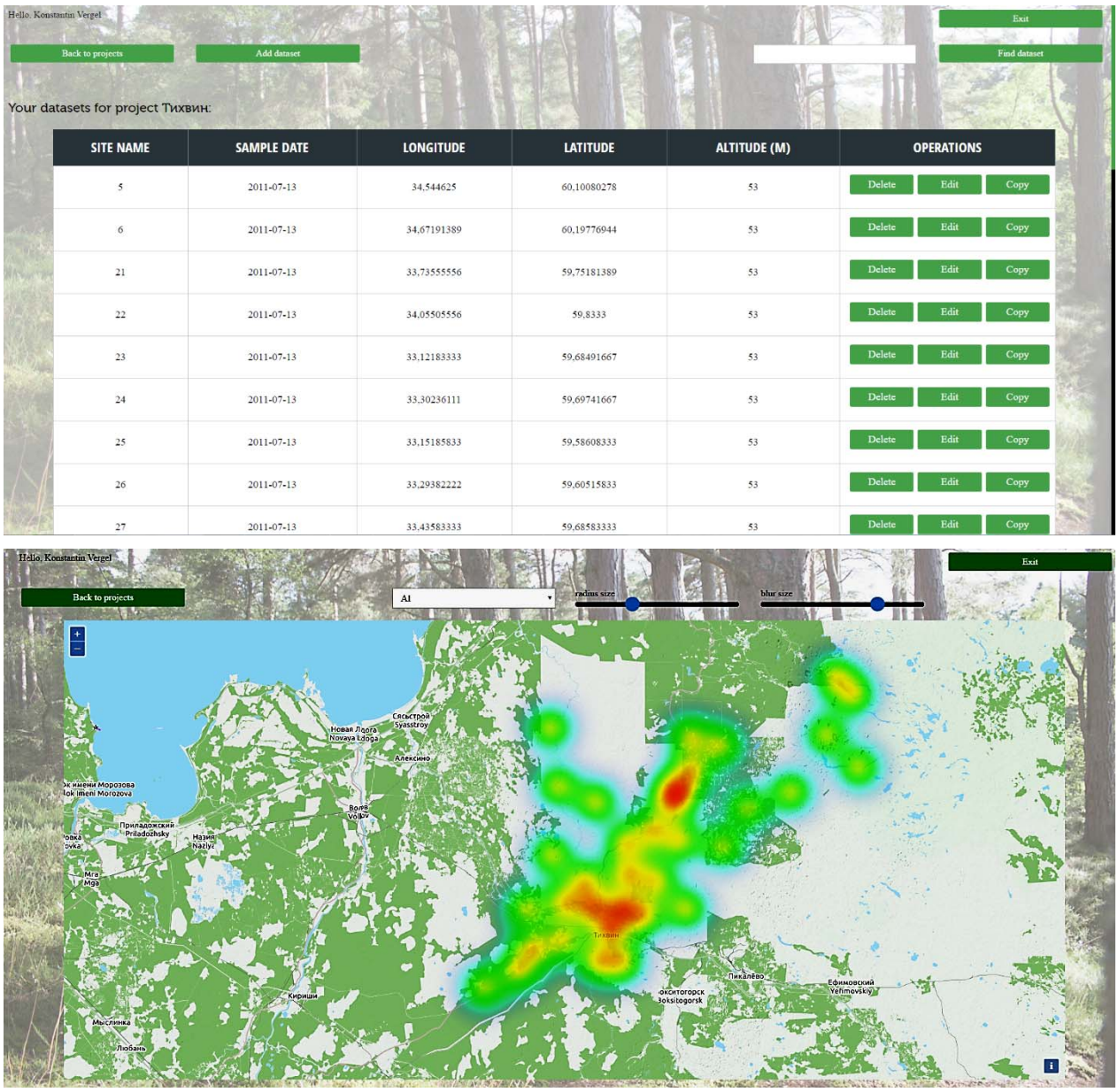


Figure 4 Web-portal interfaces

Formally, if compared to "classic" interpolation where predictor variables are limited by Cartesian coordinates, in this case a set of predictors is to be expanded with other predictor variables $g_1(x,y)$, $g_2(x,y)$,..., $g_n(x,y)$. These can be topographical features, elevation, products of aerial photography and satellite imagery and many more different surface properties.

Thus the interpolation method is replaced in this way: given

$$f(x_i,y_i,g_1(x_i,y_i),g_2(x_i,y_i),\ldots,g_n(x_i,y_i)) = f(x_i,y_i)$$
$$\forall (x_i,y_i) \in V,$$

build a predictor

$$\hat{f}(x,y,g_1(x,y),g_2(x,y),\ldots,g_n(x,y))$$

and establish an equality

$$\hat{f}(x,y) = \hat{f}(x,y,g_1(x,y),g_2(x,y),\ldots,g_n(x,y)).$$

While extended form for $\hat{f}(x,y)$ seems more complicated, it simply allows an interpolation method to fuse in possible non-linear correlations and make "better" predictions, while other formal parts of the method stay the same.

A modification of kriging called cokriging has been proposed [16]. It is oriented on using these secondary variables, as aerial and satellite protography, but has some problems with applying them to real-world data. Kriging is oriented on data that is normally distributed. While it is somewhat true for lags of spatial coordinates that are utilized in semivariogram and covariance function, it is not always true for secondary predictors g_i. Also, it is problematic to construct a covariance function as it may naturally be anisotropic and even non-symmetric. An artificial neural network, on the other hand, automatically adapts to nonlinearities and non-normally distributed variables.

This approach also constitutes some problems which are inherent to artificial neural networks. As each concrete neural network is a product of a learning

procedure, some form of predictor evaluation has to be incorporated. Usually, several different learning procedures and network topologies are evaluated and results of interpolation procedure are analyzed for deficiencies like overfitting. Such superprocedure effectively increases computational costs and may be sped up with parallel computing. Other problems are caused by data specifics, so some approaches of regularization should be employed, like learning with Gaussian noise.

Different types of satellite imagery are currently employed in data processing, like LandSat [17] and MODIS (the Moderate-resolution imaging spectroradiometer [18]). The latter project incorporates two satellites with spectroradiometers (hence the name) that is able to take satellite imagery with high spectral resolution of 36 spectral bands. Whilst, if compared to LandSat, it has moderate resolution (hence the name), it allows deeper and more thorough analyses of Earth surface, thus enabling interesting possibilities for research towards correlation and causality (e.g. contamination spreading catalysts and accelerants).

Using raw spectral radiation bands of spectral imagery is confronted with obvious interfering factors such as sun azimuth, time of day and season, surface altitude and slope and other.

Thus, in addition to the running standard statistical procedures which calculated descriptive statistics and factor analysis, neural network data processing is considered to be used in the given project, together with various MODIS products, as surface reflectance, land surface temperature, land cover, vegetation indices, land use, etc.

7 Conclusion

The study of migration and accumulation of highly toxic pollutants, which include heavy metals, persistent organic pollutants and radionuclides, the influence of pollutants on the various components of the natural and urban ecosystems is the key problem of modern biogeochemistry and ecology. The aim of the given project is to create cloud platform using modern analytical, statistical, programmatic and organizational methods to provide the scientific community with unified system of collecting, analyzing and processing of biological monitoring data.

Parts of the project have already been implemented. The rest is going to be implemented in the next two years.

References

[1] United Nations Economic Commission for Europe International Cooperative Programme on Effects of Air Pollution on Natural Vegetation and Crops (http://icpvegetation.ceh.ac.uk/

[2] Harmens H. and Mills G. (Eds.) Air Pollution: Deposition to and impacts on vegetation in (South-East Europe, Caucasus, Central Asia (EECCA/SEE) and South-East Asia. Report prepared by ICP Vegetation, March 2014. ICP Vegetation Programme Coordination Centre, Centre for Ecology and Hydrology, Bangor, UK. ISBN: 978-1-906698-48-5, 2014, 72p.

[3] Harmens H., Norris D.A., Sharps K., Mills G. … Frontasyeva M., et al. Heavy metal and nitrogen concentrations in mosses are declining across Europe whilst some “hotspots” remain in 2010. Environmental Pollution. 2015, 200:p. 93-104. http://dx.doi.org/10.1016/j.envpol.2015.01.036

[4] HEAVY METALS, NITROGEN AND POPs IN EUROPEAN MOSSES: 2015 SURVEY http://icpvegetation.ceh.ac.uk/publications/documents/MossmonitoringMANUAL-2015-17.07.14.pdf

[5] Buse A. et al. (2003). Heavy metals in European mosses: 2000/2001 survey. UNECE ICP Vegetation Coordination Centre, Centre for Ecology and Hydrology, Bangor, UK. http://icpvegetation.ceh.ac.uk.

[6] J. Alijagić, 2013. Application of multivariate statistical methods and artificial neural network for separation natural background and influence of mining and metallurgy activities on distribution of chemical elements in the Stavnja valley (Bosnia and Herzegovina) : PhD thesis. University of Nova Gorica.

[7] Žibret, G., Šajn, R., 2010. Hunting for Geochemical Associations of Elements: Factor Analysis and Self-Organising Maps. *Mathematical Geosciences*, 42(6): 681–703, doi:10.1007/s11004-010-9288-3. http://link.springer.com/article/10.1007/s11004-010-9288-3

[8] Kutovskiy N., Korenkov V., Balashov N., Baranov A., Semenov R. JINR cloud infrastructure. Procedia Computer Science, ISSN: 1877-0509, Publisher: Elsevier. 2015, 66, p. 574-583.

[9] MongoDB site and description, URL: https://www.mongodb.com/

[10] Nginx for Windows URL: http://nginx.org/ru/docs/windows.html

[11] QGIS description URL: http://www.qgis.org/en/site/

[12] OpenLayers description URL: http://docs.openlayers.org/

[13] RESTful Web services: The basics URL: https://www.ibm.com/developerworks/library/ws-restful/

[14] Goodchild M.F., Parks B.O., Steyaret L.T. 1993. Environmental modelling with GIS, Oxford University Press, New York, 488 p.

[15] S. Nickel et al. / Atmospheric Environment 99 (2014) 85e93

[16] H. Wackernagel Cokriging versus kriging in regionalized multivariate data analysis. Geoderma, 62 (1994) 83-92

[17] LandSat program description URL: http://yceo.yale.edu/what-landsat-program

[18] MODIS spectrometer on TERRA satellite URL: http://modis.gsfc.nasa.gov/about/

Ontological Approach to the Systematization of Scientific Information on Active Seismology

© Ludmila Braginskaya © Andrey Grigoryuk © Galina Zagorulko
© Valery Kovalevsky
ICM&MG SB RAS IIS SB RAS
Novosibirsk
ludmila@opg.sscc.ru and@opg.sscc.ru gal@iis.nsk.ru
kovalevsky@sscc.ru

Abstract

In this paper the principles of a knowledge portal on active seismology are discussed. The portal provides a holistic view of the subject and the various aspects of scientific activity in active seismology. A domain ontology "Active Seismology", developed by the authors was used as the conceptual framework and the information model for the portal.

1 Introduction

In the modern world, research activities directly depend on efficient access to the common digital scientific data and modern information instruments to store, search for, visualize data, and provide high level of data analysis.

The digital data resulting from observations or computational experiments in one branch of the science and presented in the form of databases, electronic collections, multi-dimensional digital arrays, etc., may be the object of a study in other disciplines. Introducing new data into the existing data archives may drastically change the structure of the scientific research object, and applying modern analysis methods to the old data may greatly increase the reliability of the research results.

In the scientific activity the researcher often has to reuse the experimental and computational data, methodology and technology of data processing and analysis, and also needs efficient access to the available publications, horizontal and vertical links between the various organizations, archiving of information, and notification services.

The purpose of the work described in the present paper is to create a scientific information system (SIS) capable of providing solutions to the above-noted problems of research in active seismology.

The portal (http://opg.sscc.ru/portal/) combines the available resources into a unified information space within the SIS and provides a reliable access to them via the Internet. A domain ontology called "Active Seismology" developed by the authors is an information model of the portal. The Internet portal is constructed with a technology [1] developed for experts in the subject areas.

This work was supported by RFBR grants no. 15-07-06821-a, no. 14-07-00832-a and the Grant of Program of the Presidium of RAS no. 18.

2 "Active Seismology" and its ontology

In classical seismology, earthquakes are considered both as an object and as an instrument of research. The in-depth studies of the Earth's crust started in the 1950 s. In the USA and Western Europe they originated from earthquake seismology and with its methods, in particular, sparse point observations per- formed with a small number of seismic stations and sometimes even small series of special explosions. In the USSR, deep seismic sounding (DSS) formed on the basis of seismic prospecting was widely used. The method of DSS was created by Academician G.A. Gamburtsev, who introduced into seismology the correlation principles of identification and tracing of reflected and refracted waves previously used in seismic prospecting. The main difference between seismology and seismic prospecting is in the use of more detailed information in the observation systems. In seismic prospecting, observation systems with multiple overlapping are widely used. Since the 1960 s, the method of continuous radiation from mechanical vibrators has been used in oil seismic prospecting. In the 1980 s, powerful seismic vibrators were created jointly by several institutes of the Siberian Branch of the USSR Academy of Sciences. With accumulation systems of recording, such vibrators create a wave effect that is equivalent to a moderate earthquake.

The sources of seismic signals are not only high-power technical facilities. They create a new scientific direction in geophysics active seismology. Active seismology provides new methods for more exact and more ecologically safe solutions of the fundamental problems of seismology of the Earth's crust structure and search for earthquake precursors [2]. The modern methods for the detection and control of variations in the parameters of elastic media in time, which are more

Proceedings of the XVIII International Conference «Data Analytics and Management in Data Intensive Domains» (DAMDID/RCDL'2016), Ershovo, Russia, October 11 - 14, 2016

sensitive than those in source seismology, have made it possible to formulate a large class of new problems. These are: monitoring of seismic prone zones, estimating the stressed state and anisotropy of the Earth's crust and upper mantle, increasing the productivity of oil formations, controlling the seismic resistance of engineering constructions, etc.

This direction of research was formed within the framework of a scientific program on Vibrational Earths Sounding performed in the Siberian Branch of the Russian Academy of Sciences in 1970 s–1990 s under the leadership of academician A.S. Alekseev. In those years, an experimental basis of the method, namely, powerful seismic vibrators, systems for recording of vibrational signals, field computational complexes, and systems for computer processing of vibroseismic data, were created. Numerous experiments in various regions of Russia were performed. During the last three decades, extensive works on active seismology have been carried out in Russia, Japan, China, USA, and some European countries.

These theoretical and experimental investigations in the field of active seismology provided considerable accumulation of information on all components of the method, including questions of the theory, creation of controllable sources, results of experimental works, and methods of mathematical simulation. This is due to the fact that the modern high-precision scientific instruments used in natural experiments generate larger and larger volumes of data. Large volumes of synthetic data have been obtained in the numerical simulation of wave processes in complex media. The problem of providing easy access to the data in the subject area of active seismology is of great importance, since large volumes of experimental data and investigation results presented in the Internet are located on various sites of scientific institutes, conferences, in electronic libraries, etc. The primary purpose of the SIS on Active Seismology is to receive, integrate, and present data and knowledge of fundamental research in the subject area. Various ontologies form the conceptual basis for the systematization of knowledge and information in the SIS.

According to [3], the ontology of knowledge area in the active seismology area is constructed on the basis of two basic ontologies, namely, ontology of research activities and ontology of scientific knowledge. It is created by means of their completion and development, which greatly simplifies both the creation of ontology, and further support. Ontology contains concepts of the modeled area, linking their relationships, attributes of concepts and relationships and restrictions on attribute values. Used for the construction of an active seismic technology [1] includes ontology language for describing ontologies, methods of construction and development of ontologies, and ontology editor.

The ontology of research activities includes a set of concepts (notions, classes) related to research activities in the field of active seismology, such as Person, Organization, Event, Activity, Publication. This set of concepts is used to describe the participants of research activities, events, projects, and various publications. Field works and field experiments are specific classes of the ontology of scientific activities in active seismology. In addition to the set of concepts, the ontology also has a set of binary relations between the concepts and a set of specimens of the classes, that is, data records corresponding to a class or relation. For instance, the class field works has a specimen called Complex Ecological-Geophysical Expedition 2005, the class field experiments, Vibrational Sounding of the Shugo Mud Volcano, and the class Person, Khairetdinov Marat Samatovich. Specimens of the class Information Systems are thematic and computational Internet resources, in particular, computational systems, databases, an electronic library, which are structural elements of the scientific information system Active Seismology [4].

The second basic ontology, the ontology of scientific knowledge, contains some meta-concepts to describe the concepts of knowledge domains in active seismology.

The construction of hierarchies of concepts is an important stage in the development of the subject area ontology. The subject area ontology on active seismology developed by the authors includes five basic hierarchies, namely, a hierarchy of branches of science, a hierarchy of objects, a hierarchy of subjects of research, a hierarchy of investigation methods, and that of scientific results.

In addition to the above-listed hierarchies of classes, which have been constructed on the basis of meta-concepts of the ontology of scientific knowledge, the subject area ontology was extended by a hierarchy of classes called Facilities including classes Source and Sensor. The class Source has three subclasses. The subclass Vibroseismic Source provides technical characteristics and locations of seismic vibrators (generators of seismic waves).

The subclass Seismic Noise is used to describe experimental works on the recording of natural seismic noise. This class is created, since in recent years the methods of active seismology have also been extended by the experimental works in which the seismic field of non-controllable natural sources is recorded. The technology of recording has been created for a specific task of geophysics, for instance, seismic emission tomography of volcanic structures using seismic noise from an active volcano zone. In 2010, a unique experiment on the recording of seismic noise by the seismic antenna method in an adit of the Baksan Neutrino Observatory of the Institute for Nuclear research was performed by the Institute of Computational Mathematics and Mathematical Geophysics SB RAS.

A subclass called Explosion describes the parameters of industrial explosions recorded during some experimental works. The experimental works on the recording of parameters of industrial explosions are performed within the frame- work of a project on vibroseismic methods to estimate the ecological risks of industrial explosions [5].

The divisions of science on Active Seismology have been created by the authors by analyzing sites of various thematic conferences, electronic libraries of scientific

organizations in this subject area, and monographs [6, 7], on active seismology. The hierarchy of concepts has been constructed with the help of a group of experts in various scientific branches of active seismology (Fig. 1).

Since the concepts of active seismology, earthquake seismology, and seismic prospecting are mostly the same, the hierarchy of subjects and methods is based on a classification proposed by Academician N.N. Puzyrev [8]. This classification makes it possible to describe the modern seismic methods of studying the structure of Earth's interior, including search for minerals, and spontaneous seismic processes (earthquakes) from a unified point of view. The ontology of hierarchies of subjects and methods of research proposed by the authors has been revised and extended by concepts that are specific for the domain of active seismology (Fig. 2).

In this ontology, scientific results mean experimental and theoretical data on active seismology, as well as results of analysis of these data. In addition to the class-subclass interrelations, the ontology concepts may have associative interrelations. The most important associative interrelations between the concepts of research activities and scientific knowledge are: investigate (compare research activities or a division of science with the object of investigation); use (link the investigation method with a type of activity, researcher, or a division of science); apply to (link the investigation method with the object of investigation); describe (link a publication with a scientific result, object, or method of investigation). An important class, which describes the databases and computer systems on active seismology, is called Information Resources.

From a substantive point of view an ontology of active seismology, built by expanding the two basic ontologies, can serve as a representation of concepts for active seismology and vibroseismic monitoring, and carried out by persons and organizations activities in the framework of this research area.

3 Information system and data portal on active seismology

The scientific information system (SIS) Active Seismology (http://opg.sscc.ru) is designed to support the theoretical and experimental investigations in the subject area. Its main components are: an information computational system (ICS) [9] called Vibroseismic Earths Sounding, which provides the users with multi-parametric search, computing-analytical, and GIS services for online work with data of vibroseismic monitoring; a database of scientific works, which can be augmented by the users, an electronic library; and a bibliographic catalog also augmented by the users.

The information basis of the data portal is the ontology, which combines the resources into a unified information space and provides access to them via the Internet. The Internet portal is used to combine diverse information into the unified information space, provide easy access to it, and efficiently control the content. The domain ontology Active Seismology is used as the conceptual basis and information model of the data portal.

The knowledge portal on active seismology (http://opg.sscc.ru/portal/) is designed to systematize this subject area on the whole and the diverse data and processing facilities in the SIS.

The ontology described above is used as the conceptual basis of the information model of the knowledge portal. The ontology of the portal contains formal descriptions of concepts in the subject area in the form of classes of objects and interrelations between them, thereby providing structures to represent real objects and interrelations between them. Accordingly, data on the portal are presented as a semantic network, that is, as a set of diverse interrelated information objects. Comprehensive access to the systematized knowledge and information resources is provided by well-developed navigation and search facilities whose functioning is also based on the ontology.

Currently, the portal provides a reduction in the common information space only Russian-language information resources. portal interface is implemented in Russian.

Figure 3 shows a page with a description of an object called "Elbrus Experiment", with the corresponding objects represented by hyperlinks. These hyperlinks provide a description of the organizations and personnel engaged in this activity, investigation methods used, and access to the experimental database and the information computer system.

Figure 4 displays detailed information (metadata) about the experiment by a user's web-browser: coordinates of stations for seismic signal recording, recording equipment, schedule of recording sessions, experimental map, and results of analysis of seismic signals. According to the enquiry shown in the figure, wave forms and frequency-time characteristics of seismic events corresponding to 1830–2000 for sessions no. 11 and no.12 are presented.

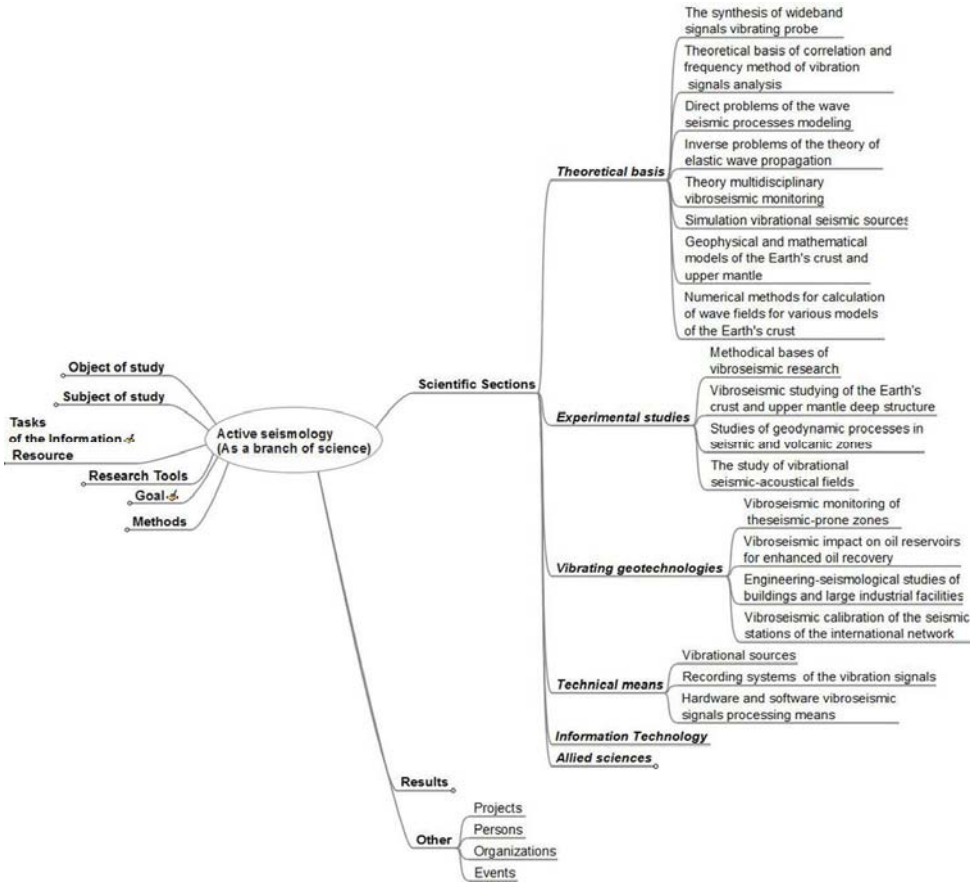


Figure 1 Hierarchy of concepts of active seismology (a fragment)

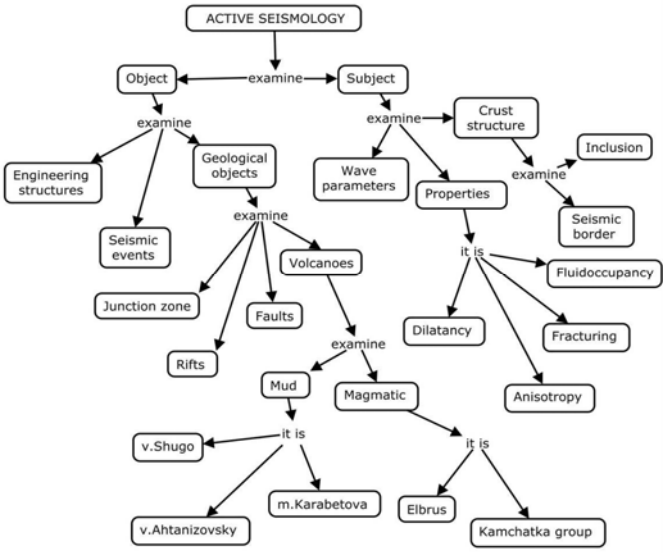
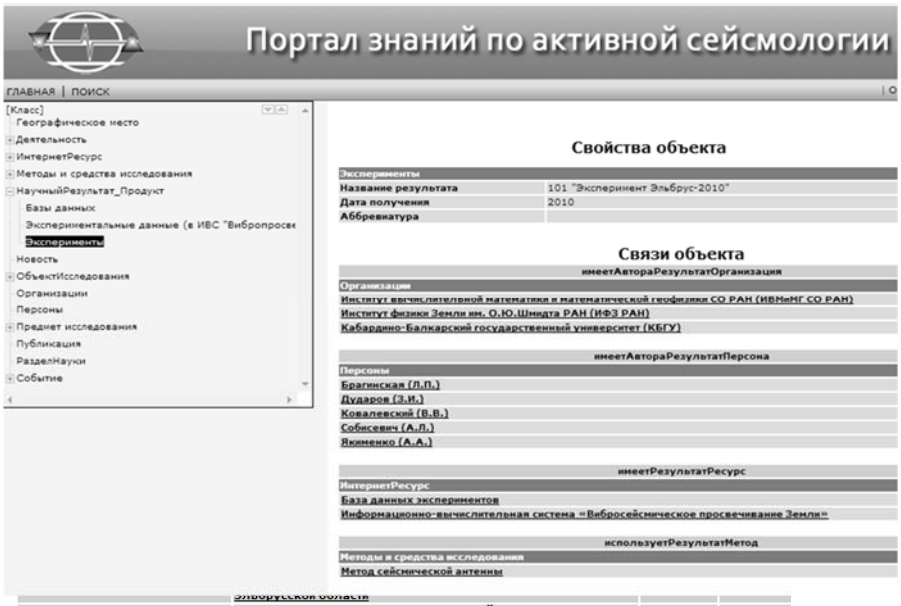


Figure 2 Ontology (a fragment)



Собисевич (А.Л.)	ОТЧЕТ О НАУЧНО-ИССЛЕДОВАТЕЛЬСКОЙ РАБОТЕ по теме: Исследование геодинамических процессов в зоне возникновения ожидаемых землетрясений на Северном Кавказе с использованием уникальной установки «Комплексная геофизическая информационно-измерительная система Кабардино-алкарского государственного университета им. Х.М. Бербекова	2009	
Брагинская (Л.П.) Дударов (З.И.) Ковалевский (В.В.) Собисевич (А.Л.) Якименко (А.А.)	Отчет по экспедиционным работам в Баксанской штольне	2010	отчет
Масуренков (Ю.П.)	Пульсационно-вихревое развитие Эльбрусской вулканической области (как следствие миграции мантийного плюма?).	2010	статья
Масуренков (Ю.П.)	Современная деятельность вулкана Эльбрус. Доклады АН СССР, т. 142, № 6, 1962.	1962	статья
Масуренков (Ю.П.)	Современное состояние вулкана Эльбрус	1971	статья
Масуренков (Ю.П.)	Тектоника, магматизм и углекислые воды Приэльбрусья	1961	статья
Масуренков (Ю.П.)	Эльбрусский магматический очаг	1964	статья

Figure 3 Description, map, and display of the recorded seismic events and results of analysis of Elbrus-2010 Experiment in the SIS Active Seismology

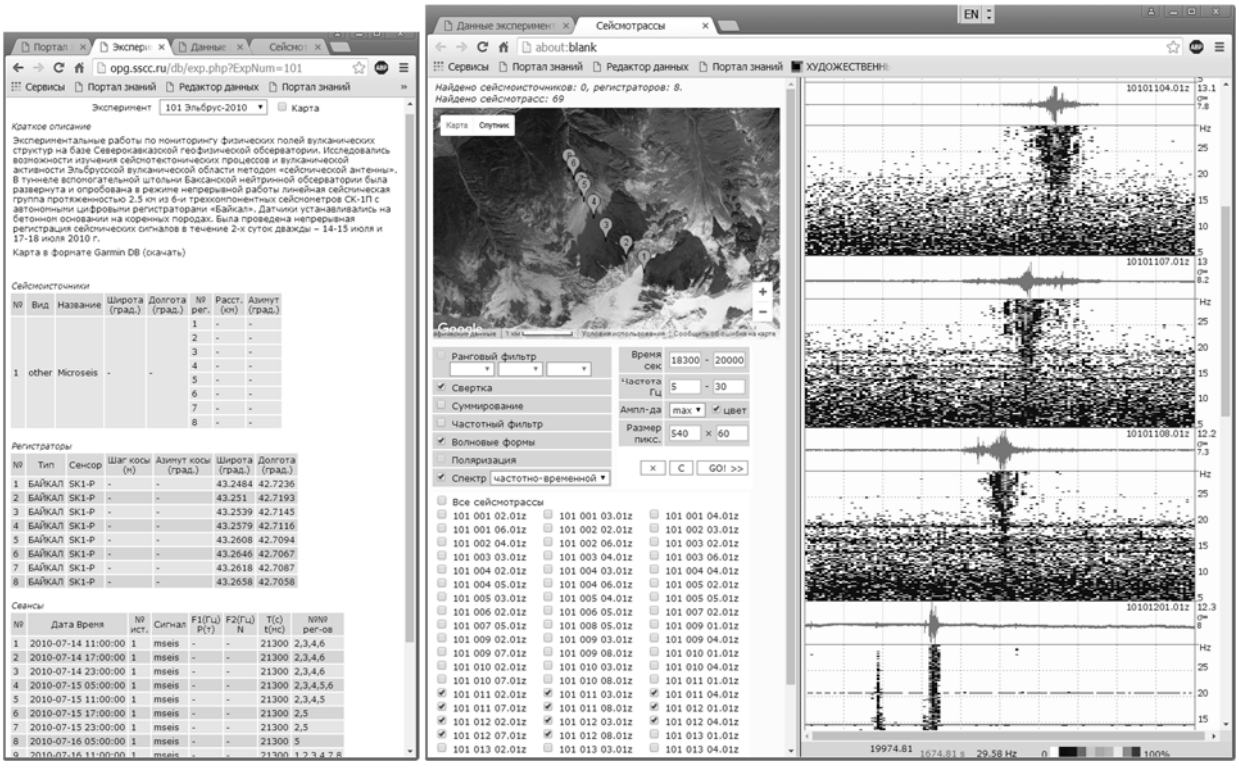


Figure 4 Description, map, and display of the recorded seismic events and results of analysis of "Elbrus-2010 Experiment" in the SIS "Active Seismology"

4 Conclusions

In this paper, a scientific information system to systematize the knowledge and data on active seismology was described. The following results have been obtained:

1. The available data on Active Seismology have been systematized on the basis of ontology within the framework of a unified information space. The ontological approach widely used now to systematize knowledge in many subject areas is used for the first time in the field of active seismology.

2. A user friendly access has been created not only to the information about the branch of science under consideration, but also to the available field data obtained in long-term experiments on Earths vibroseismic monitoring.

The SIS Active Seismology can be used by scientists and engineers employing the methods and theoretical and experimental data on active seismology, by students specializing in this field, and by specialists estimating the state of the art in this scientific area of research.

References

[1] Yury Zagorulko, Olesya Borovikova, Galina Zagorulko. Knowledge Portal on Computational Linguistics: Content-Based Multilingual Access to Linguistic Information Resources. Selected topics in Applied Computer Science. In Proceedings of the 10th WSEAS International Conference on Applied Computer Science (ACS10). Hamido Fujita, Jun Sasaki (Eds.). (Iwate Prefectural University, Japan, October 4–6, 2010). WSEAS Press, 2010. p. 255–262.

[2] Alekseev A.S., Glinskii B.M., Kovalevskii V.V. et al. Active Seismology with Powerful Vibrational Sources. Editor in Chief G.M. Tsibulchik,. Novosibirsk: Inst. of Comp. Math. and Math. Geoph. Publ., GEO SB RAS Publ., 2004.

[3] Yury Zagorulko, Galina Zagorulko. Ontology-Based Technology for Development of Intelligent Scientific Internet Resources. Intelligent Software Methodologies, Tools and Techniques. In Proceedings of the 14th International Conference, SoMet 2015, Naples, Italy, September 15-17, 2015. Proceedings. Hamido Fujita, Guido Guizzi (Eds.), Communications in Computer and Information Science, Vol. 532, Springer International Publishing Switzerland 2015. p. 227–241.

[4] Kovalevsky V.V., Braginskaya L.P., Grigoryuk A.P. Experimental data management using modern web technologies. The BSU Bulletin, 2, February 2013 http://www.bsu.ru/content/page/1466/2_2013.pdf

[5] Informative factors of geophysical fields interaction in problems of the environmental protection prediction M. S. Khairetdinov, V. V. Gubarev, G. M. Voskoboynikova, G. F. Sedukhina. In Proceedings of the International Conference Computational and Informational Technologies in Science, Engineering and Education (CITech 2015), Kazakhstan, Almaty, 2427 Sept. 2015. Almaty : Kasak ., 2015. p. 128

[6] A.S. Alekseev, G.M. Tsibulchik, V.V. Kovalevsky, A.S. Belonosov. The basis of the theory of active geophysical (multidisciplinary) monitoring. Conception of source and surface dilatant zones. Handbook of Geophysical Exploration: Seismic Exploration, Active geophysical monitoring, 2010, Vol. 40, p. 105–133

[7] A.S. Alekseev, B.M. Glinsky, V.V. Kovalevsky, M.S. Khairetdinov, 2010, Active vibromonitoring: experimental systems and fieldwork results. Handbook of Geophysical Exploration: Seismic Exploration, Active geophysical monitoring, 2010, Vol. 40, p. 55–71

[8] N.N. Puzyrev. History of multiwave seismic exploration in Russia. Geology and Geophysics, 2003, 44(4), p. 277–286

[9] Grigoryuk A.P., Kratov S.V. Data Experiments Management on Web-Technologies Basis. In Proceedings of the 12th International Conference on Actual Problems of Electronic Instrument Engineering, APEIE 2014, Novosibirsk, October 2014, Vol. 3, p. 259–261.

Инфраструктуры данных в астрономии

Data Infrastructures in Astronomy

Звездные скопления: развитие знаний на основе интенсивного использования данных

© С.В. Верещагин, © Н.В. Чупина, © А.С. Фионов

Институт астрономии РАН
Москва
svvs@ya.ru chupina@inasan.ru fionov@mail.ru

Аннотация

Представлены результаты анализа информации о рассеянных звездных скоплениях (РЗС). Изучены динамика роста числа публикаций по РЗС, рост объема каталогов РЗС и звездных обзоров. Показано, что применение фильтров позволит обнаружить более 10 тысяч новых РЗС. Поставлена задача использования методов и инструментов интенсивного использования данных для работы со структурированной и неструктурированной информацией о РЗС.

1 Введение

На примере информационного обеспечения исследований рассеянных звездных скоплений рассмотрены особенности интенсивного использования данных в астрономии (см. также [20]). Цель работы – подготовка требований к разработке методов автоматического поиска и изучения звездных скоплений путем оперативного использования наиболее полной информации о РЗС. Объединение неструктурированной информации, содержащейся в публикациях, со структурированной информацией в обзорах звезд и каталогах РЗС, поиск ранее неизвестных скоплений и индексация звезд, входящих в их состав, позволят получить новые результаты об устройстве и кинематике, как диска Галактики, так и отдельных РЗС. Например, изучение морфологии области в Орионе позволило найти новую группу звезд [3]. Обнаружены и изучаются группировки звезд внутри скопления M67 [2]. Эти работы показали важность индексации звезд для верификации результатов путем подключения дополнительных параметров этих звезд из каталогов и публикаций (фотометрия, параллаксы, химический состав) и для поиска новых звезд, входящих в состав этих группировок.

В [9] (и других работах этих авторов) представлена информационная система, позволяющая осуществлять поиск новых РЗС. Система включает информационный блок, состоящий из структурированной информации о каждом из скоплений, звездах в составе скопления и неструктурированную информацию, такую, как диаграммы и карты. Таким образом, существует тенденция перехода от простых файлов-каталогов к информационным системам, содержащим многообразную информацию об изучаемых объектах. Новый обзор GAIA (Global Astrometric Interferometer for Astrophysics) [15] позволит продолжить эту работу в направлениях поиска новых и изучения известных РЗС. Ниже сделана оценка ожидаемого числа открытий скоплений в ближайшем будущем и выполнен анализ роста числа публикаций по теме РЗС при помощи Astrophysical Data System [1]. Обсуждение и выводы содержат формулировки требований к возможности применения инструментов интенсивного использования данных к разнородной информации о РЗС.

2 Рассеянные звездные скопления

РЗС представляют собой группировки с численностью звезд от десятков до тысяч в их составе, [10]. На небе скопления выглядят как нерегулярные группы различных размеров, часто с заметным повышением концентрации звезд к центру. Внутри скопления звезды связаны гравитационно, часть из них уже покинула скопление путем диссипации. Звезды в скоплении имеют примерно одинаковые возрасты, хотя существуют скопления с продолжающимся звездообразованием и по этой причине могут содержать молодые звезды. Несмотря на обилие данных наблюдений и множество публикаций, внутреннее устройство РЗС и система скоплений в диске изучены недостаточно. Для РЗС наиболее надежно определены расстояния от Солнца, возрасты, химические составы и т. п. Располагаясь на различных расстояниях от Солнца, они являются ключевыми объектами для получения новых результатов, важных для понимания деталей строения Галактики. Особое место занимают исследования внутреннего устройства и кинематики скоплений. Для этого важно пополнять информацию об отдельных звездах в составе скоплений путем индексации и применения инструментов интенсивного использования данных.

Большинство РЗС образовалось очень давно, но молодых скоплений наблюдается больше, чем

Труды XVIII Международной конференции DAMDID/RCDL'2016 «Аналитика и управление данными в областях с интенсивным использованием данных», Ершово, 11-14 октября 2016

старых. Такое распределение отражает факт распада скоплений со временем: взаимодействуя друг с другом, некоторые звёзды приобретают скорости, большие, чем скорость отрыва, и покидают скопление. Молодые рассеянные скопления расположены вблизи плоскости Галактики на высоте не превосходящей 200 пк, что близко к характеристикам распределения поглощающей материи. В большей степени отходят от плоскости Галактики наиболее старые скопления, некоторые из них достигают высоты до 2.2 кпк. Для старых скоплений не заметна концентрация к плоскости Галактики. Возможно, быстрее разрушаются те скопления, которые движутся вблизи галактической плоскости. Исследования здесь далеки от завершения – альтернативная гипотеза состоит в предположении, что ранее скопления рождались в более толстом слое диска, чем в настоящее время.

Интересно, что в составе одного скоплении можно наблюдать звезды самых разных типов, кратные системы и группировки. Сосредоточение разнообразных объектов на площадке в несколько квадратных градусов на небе, занимаемой скоплением, – хорошая лаборатория для исследования звезд. Особенно удобны такие наблюдения для поиска звезд с экзопланетами.

3 Рост публикаций об РЗС

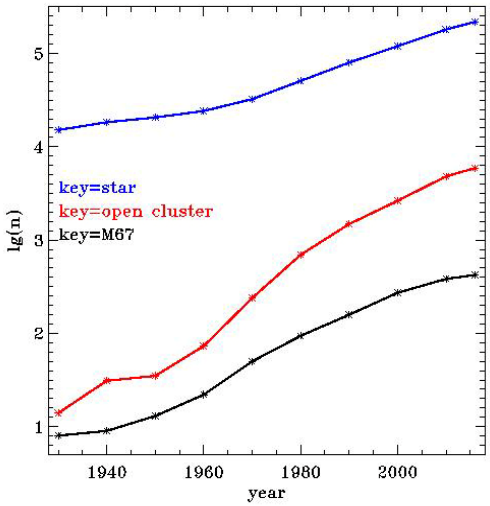
Вначале 1980-х гг. число известных РЗС, содержащихся в “карточном” каталоге Рупрехта, было около 1200. В настоящее время по данным [14] их насчитывается 3784 и это далеко не предел. Всего по приближенным оценкам в Галактике более 100 тысяч РЗС, и основные открытия еще впереди.

Таблица 1 Рост информации со временем

Дата, до (год)	**“M67”**	**“open (and) cluster”**	**“star”**	**k**
1930	8	14	15103	1079
1940	9	31	18313	590
1950	13	35	20590	588
1960	22	73	24138	330
1970	50	239	32429	136
1980	94	695	50711	73
1990	158	1483	79757	54
2000	272	2632	119757	46
2010	383	4815	180543	37
2016	422	5914	218392	37

С помощью системы ADS [1] мы оценили рост информации по РЗС, звездам и отдельному скоплению M67. В Таблице 1 приведено количество публикаций (по наличию в заголовке статьи ключей “open (and) cluster”, “star” и “M67”) до указанного года. Коэффициент k – отношение чисел в колонках “star” и “open cluster”. С его помощью показан рост популярности исследований скоплений по сравнению с исследованиями звезд: начиная с 1980-х гг. он заметно уменьшился (с ростом года в колонке “Дата”), что свидетельствует о возрастании числа запросов по ключу “open (and) cluster” относительно “star”.

На Рис. 1 видно, что информация об РЗС (ключ “open (and) cluster”) растет опережающим темпом по сравнению с более общей темой про звезды (ключи “star” и “M67”). Отметим увеличения скорости роста числа статей по РЗС в 1950-е годы и по звездам в 1970-е. Для РЗС это было связано, скорее всего, с резким увеличением числа наблюдений. Для звезд с 1970-х годов появление крупных ЭВМ позволило резко увеличить исследования по их физике и эволюции. Отметим, что число публикаций по звездам значительно превосходит число публикаций по РЗС. Это вполне естественно, если учесть огромное превосходство в численности и разнообразии звезд в сравнении с РЗС. Сегодня наметился спад публикаций по РЗС, что мало заметно и для звезд. Для M67 рост также замедляется, что иллюстрируют кривые на Рис. 1.



- Логарифм числа статей (n) с ключами в заголовках статей “star”, “open (and) cluster”, “M67”, вышедших до указанной даты (year)

4 Звездные обзоры

Поиски и открытия новых РЗС происходят в основном путем применения различных фильтров к распределениям звезд из современных обзоров, краткая информация о некоторых наиболее крупных из них представлена в Таблице 2. В Таблице 2 представлены поверхностные звездные плотности (P), численности звезд, процент охвата неба и предельные звездные величины рассмотренных обзоров. Обзор GAIA [15] включает приблизительно 1 млрд. звезд, что составляет приблизительно 1% звездного населения Млечного Пути. 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 кинематике и строении РЗС. Точность достигнута во многом благодаря сверхдальнему (более 1 млн. км от Земли) расположению ИСЗ в точке Лагранжа (L2),

исключающей влияние на положение аппарата Земли–Луны и Солнца. Исключаются засветка от Земли, Луны (отраженного света от Земли), а также влияние переходов из света в тень. Последние мешают работе близких аппаратов, таких, как телескоп Хаббл. Кроме того, поддержание аппарата в точке L2 наиболее энергетически выгодно [7].

Таблица 2 Поверхностные звездные плотности в звездных обзорах

Обзор	Число звезд	Охват неба, %	P, зв./ кв. град.	Пред. зв. вел.
GAIA	10**9	100	24240	G=20
UCAC4	0.113*10**9	100	2739	R=16
2MASS	0.300*10**9	70	4848	J=15.8

UCAC4 (United States Naval Observatory CCD Astrograph Catalog) [19] представляет собой компилятивный каталог, включающий 113,780,093 звезд главным образом на интервале от 8 до 16 зв. величины в полосе пропускания между V и R. Ошибки положений составляют от 15 до 20 мкс для звезд в диапазоне от 10 до 14 зв. величин. Получены собственные движения для большинства звезд. Для этого использовались около 140 звездных каталогов с существенным различием эпох наблюдений. Эти данные дополнены фотометрией из 2MASS приблизительно для 110 миллионов звезд. В частности, 5- полосная (B, V, г, г, I) фотометрия подключена более чем для 50 миллионов звезд [6]. Обзор UCAC4 полон для ярких звезд (приблизительно до R = 16 зв. величины). По номерам звезд обеспечена связь с каталогом Hipparcos [18], где есть дополнительные данные, например, параллаксы.

2MASS (Two Micron All-Sky Survey) [17] – это обзор неба в ИК диапазоне в полосах пропускания J (1.25 μm), H (1.65 μm) и Ks (2.17 μm). Осуществлен в 1997–2001 годах с помощью телескопов в северном (Обсерватория имени Уиппла, Аризона) и южном (Чили) полушариях. Каталог точечных источников содержит положения и потоки для приблизительно 300 млн. звезд до J=15.8.

5 Фильтры для поиска РЗС

Применение фильтров к звездным обзорам позволяет многократно увеличить число открытых РЗС. Большие перспективы открываются в ближайшем будущем, как уже говорилось, с поэтапной реализацией проекта GAIA. Первые результаты появятся в общем доступе в сентябре 2016 года. Доступные данные будут постепенно увеличиваться в ближайшие годы [15]. Объем обзора достигает 200 Тб.

О методах выделения пространственных флуктуаций звездной плотности (фильтрах) см. [13]. Для выявления мест повышенной концентрации звезд применяются методы “ближайшего соседа”, детализации полигональной сетки Вороного, нахождения минимального остовного дерева. Чтобы понять, являются ли выделенные уплотнения звезд реальными РЗС, применяются кинематический, фотометрический, статистический и другие критерии.

Кинематический критерий [4] основан на методе диаграмм апексов (AD-диаграмм), разработанном авторами. Он применялся нами для исследования кинематики Гиад, Яслей, скоплений и групп в Орионе, потока Большой Медведицы. AD-диаграмма представляет собой распределение апексов звезд в экваториальной системе координат. Их координаты 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 можно назвать индивидуальными апексами звезд. Как отмечается в [22], РЗС обнаруживались даже с помощью простого визуального просмотра распределений звезд – по данным обзоров DSS и 2MASS в [12] были обнаружены 66 кандидатов в скопления. В работе [6] путем автоматического поиска флуктуаций плотности по картам распределения звездной плотности, построенных по данным 2MASS, обнаружено 681 скопление. В [11] разработан новый эффективный метод поиска пиков звёздной плотности на звездных картах, построенных по многим (79) обзорам. Этот метод основан на свертке карт плотности со специальным двумерным фильтром.

6 Сравнение поверхностных плотностей

Каково же число потенциальных открытий РЗС в обзоре GAIA? Мы сделали такую оценку путем сравнения средних плотностей звезд в скоплениях (ρ) и обзорах (P). Плотности P приведены в Таблице 2. В РЗС ρ зависит, например, от возрастов (Рис. 2 и 3) и, главным образом, от расстояния скопления от Солнца (Рис. 2). На Рис. 2 и Рис. 3 показаны эти зависимости, которые мы построили по данным каталога MWSC II [10].

Конкретный пример - скопление M67 рассмотрено особо. Мы взяли его средние параметры – площадь 3 кв. град. и численность 1000 звезд в его составе. В этом случае плотность ρ равна 318 звезд на кв. град. Как видим из Таблицы 2, все звезды M67 с большим запасом входят в рассмотренные обзоры, поскольку P любого из обзоров, как минимум на порядок, превосходит ρ для M67.

Плотность в GAIA (Таблица 2) примерно в три раза превосходит плотность максимально плотного РЗС в MWSC II (до 8000 звезд на кв. град., Рис. 2). Это означает, что в GAIA содержится по крайней мере в три раза больше РЗС, чем в [10] - 3006. Следовательно, в GAIA содержится приблизительно 10 тыс. ранее неизвестных РЗС, которые ждут первооткрывателя. Отметим, что в 2MASS и UCAC4

входят далеко не все скопления [10], что можно определить по их значениям P относительно GAIA в Таблице 2.

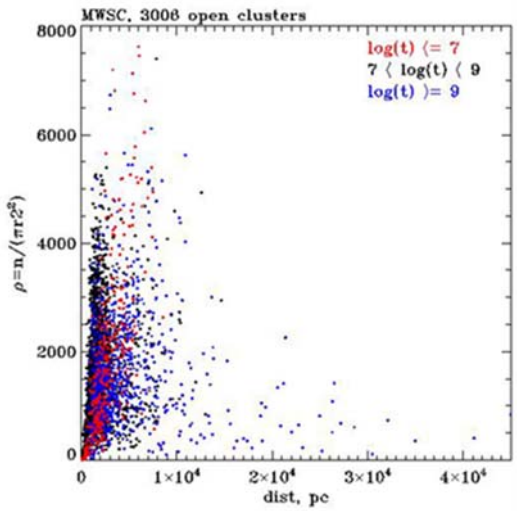


Рисунок 2 Распределение РЗС из MWSC II. По оси ординат – поверхностная плотность ρ звезд в скоплениях, по оси абсцисс – расстояние скоплений от Солнца (dist). Цветом обозначены РС разных возрастных диапазонов, подписанных в справа вверху

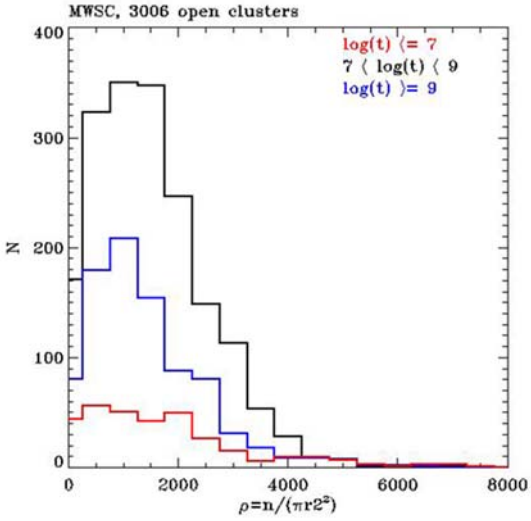


Рисунок 3 Число скоплений (N) в зависимости от ρ. Гистограммы построены отдельно для РЗС разных возрастов, обозначенных цветом и подписанных справа вверху

7 Метаданные

Для одновременной работы с каталогами РЗС, публикациями и звездными обзорами нужны метаданные. Ими являются номера РЗС в различных каталогах, их экваториальные координаты и параметры - радиус, расстояние от Солнца, химический состав и т. п.. В состав метаданных также входит номер звезды внутри скопления. В [9] предложен вариант нумерации WOCS звезд в скоплениях ID (IDW) и индексация ID (IDX). В свою очередь ID основаны на звездной величине в фильтре V и расстоянии звезды от центра скопления. В качестве метаданных целесообразно использовать и неструктурированную информацию, которая основополагающа при исследовании РЗС - диаграммы Герцшпрунга–Рессела (H-R), двуцветные диаграммы (CMD) и диаграммы собственных движений (PMD).

8 Заключение

Показано присутствие в информации о РЗС характеристик Big Data (так называемых «трех V», [8, 23]): гигантского объема (volume) постоянно растущей структурированной (обзоры, каталоги) и неструктурированной (публикации) многообразной (variety) информации. Степень изученности звездного населения Млечного Пути составляет 1% (как уже говорилось, GAIA включает данные об 1 из 100 млрд. звезд Галактики). В GAIA процент охвата РЗС составляет также около 1% всех РЗС в Галактике, предоставляя возможность открыть еще 10 тыс. скоплений.

Важным в развитии метаданных является проведение индексации звезд, входящих в состав РЗС. Это ускорит наполнение данными, поиск новых членов скоплений и другие исследования. Повышение скорости необходимо также для организации совместного поиска в нескольких обзорах (на сегодня – GAIA, 2MASS), каталогах РЗС и системе публикаций ADS (с возможностью поиска по параметрам).

Скорость прироста (velocity) составляет несколько тысяч публикаций в год, наблюдается многократное увеличение объема обзоров (Таблица 2) и числа открытий РЗС. Применение инструментов интенсивного использования данных к РЗС будет эффективным для получения всесторонней информации о выбранном скоплении. Физические параметры, результаты наблюдений, публикации и другая информация – это обычная подготовка материалов для исследований. Автоматизация с помощью инструментов интенсивного использования данных ускорит процесс сбора материалов.

Подход к объединению структурированной и неструктурированной информации об РЗС существует в CDS [16]. Здесь разработан аппарат поиска каталогов РЗС и данных о звездах без объединения поисковых возможностей. Разработана удобная система для объединения и обработки данных TOPCAT. Система работы с данными GAIA также будет удаленной, [15]. Авторами [10] создана система Каталогов рассеянных скоплений Млечного пути. Система содержит результаты поиска новых РЗС на базе каталога 2MASS, а также множество параметров скоплений, отдельно оформлены точки входа для карт и диаграмм.

Благодарности

Авторы благодарны QI и рецензентам за помощь. Работа поддержана грантом РФФИ 16-52-12027 ННИО_а.

Литература

[1] Astrophysical Data System, SAO/NASA ADS Astronomy Abstract Service http://adsabs.harvard.edu/abs/

[2] N.V. Chupina, S.V. Vereshchagin. Stellar clumps within the corona in the open cluster M 67 // Astronomy and Astrophysics, V. 334, p. 552–557, 1998.

[3] N.V. Chupina, S.V. Vereshchagin. Star clusters in the Sword region in Orion // In Proceedings of the 33rd ESLAB Symposium on Star formation from the small to the large scale, p. 347–349, edited by F. Favata, A. Kaas, and A. Wilson, ESA SP 445, Noordwijk, The Netherlands, ESA, 2000.

[4] N.V. Chupina, V.G. Reva, S.V. Vereshchagin. The geometry of stellar motions in the nucleus region of the Ursa Major kinematic group // Astronomy and Astrophysics, V. 371, p. 115–122, 2001.

[5] R.M. Cutri и др. 2MASS (The Two Micron All-Sky Survey) The 2MASS All-Sky Catalog of Point Sources University of Massachusetts and Infrared Processing and Analysis Center (IPAC/California Institute of Technology), 2003. Cat. II/246.

[6] D. Froebrich, A. Scholz, C.L. Raftery. A systematic survey for infrared star clusters with |b| <20° using 2MASS // Monthly Notices of the Royal Astronomical Society, V. 374, p. 399, 2007.

[7] Gaia enters its operational orbit, The European Space Agency (ESA) 2014-01-08.

[8] Gartner, Gartner Says Solving 'Big Data' Challenge Involves More Than Just Managing Volumes of Data STAMFORD, Conn., June 27, 2011. http://www.gartner.com/newsroom/id/1731916

[9] K. Tabetha Hole, Aaron M Geller, Robert D. Mathieu, Imants Platais, Søren Meibom, David W. Latham. WIYN Open Cluster Study. XXIV. Stellar Radial-Velocity Measurements in NGC 6819 // The Astronomical Journal, V. 138, Issue 1, p. 159–168, 2009.

[10] N.V. Kharchenko, A.E. Piskunov, E. Schilnach, S. Roeser, R. Scholz, MWSC Milky Way global survey of star clusters. II // Astronomy and Astrophysics, V. 558, p. 53, 2013.

[11] S.E. Koposov, E.V. Glushkova, I.Yu. Zolotukhin. Automated search for Galactic star clusters in large multiband surveys. I. Discovery of 15 new open clusters in the Galactic anticenter region // Astronomy and Astrophysics, V. 486. p. 771, 2008.

[12] M. Kronberger, P. Teutsch, B. Alessi, M. Steine, L. Ferrero, K. Graczewski, M. Juchert, D. Patchick, D. Riddle, J. Saloranta and 2 coauthors. New galactic open cluster candidates from DSS and 2MASS imagery // Astronomy and Astrophysics, V. 384, p. 403, 2006.

[13] S. Schmeja. Identifying star clusters in a field: a comparison of different algorithms // Astronomische Nachrichten, V. 332, No 2, p. 172–184, 2011.

[14] S. Schmeja, N.V. Kharchenko, A.E. Piskunov, S. Röser, E. Schilbach, D. Froebrich, and R. Scholz. Global survey of star clusters in the Milky Way III. 139 new open clusters at high Galactic latitudes // Astronomy and Astrophysics, V. 568, A51, 2014.

[15] Science with 1 billion objects in three dimensions – Gaia. http://www.cosmos.esa.int/web/gaia/

[16] Strasbourg astronomical Data Center. http://cds.u-strasbg.fr/

[17] The Two Micron All Sky Survey at IPAC. http://www.ipac.caltech.edu/2mass/overview/access.html

[18] The Hipparcos and Tycho Catalogues, ESA SP-1200 – Cat. I / 239.

[19] USNO CCD Astrograph Catalog (UCAC). http://www.usno.navy.mil/USNO/astrometry/optical-IR-prod/ucac

[20] N. Zacharias, C.T. Finch, T.M. Girard, A. Henden, J.L. Bartlett, D.G. Monet, M.I. Zacharias. The fourth U.S. Naval Observatory CCD Astrograph Catalog (UCAC4) // The Astronomical Journal, V. 145, p. 44, 2013. http://cdsarc.u-strasbg.fr/viz-bin/Cat?I/322A

[21] Y. Zhang, Y. Zhao. Astronomy in the Big Data Era // Data Science Journal, V. 14, p. 11, 2015. http://datascience.codata.org/articles/10.5334/dsj-2015-011/

[22] Е.В. Глушкова. Комплексное исследование рассеянных звездных скоплений Галактики // Дис. … доктора физ.-мат. наук. Москва, ГАИШ МГУ, 2014.

[23] Крис Канаракус. Машина Больших Данных // Журнал «Сети» № 04, 11.10.2011. http://www.osp.ru/nets/2011/04/13010802/

Star clusters: the growth of knowledge based on data intensive research

S.V. Vereshchagin, N.V. Chupina, A.S. Fionov

The results of information analysis on the open star clusters (OC) are presented. The growth in number of publications on OC, stellar reviews and OC catalogs has been demonstrated. It is shown that the use of filters allows to detect more than 10 thousand of unknown OC. The main intention of the paper is to show prospects of the use of methods and tools of data intensive research to analyze the structured and unstructured information on OC.

Круглосуточный радио обзор неба на 110 МГц: база данных наблюдений и статистический анализ импульсных явлений в 2012-2013 гг.

В. А. Самодуров[1,2], А.С. Позаненко[3], А.Е. Родин[1],
Д.Д. Чураков[4], Д.В. Думский[1,2], Е.А. Исаев[1,2],
А.Н. Казанцев[1], С.В. Логвиненко[1], В.В. Орешко[1],
М.О. Торопов[5], М.И. Волобуева[6]
[1]Пущинская Радиоастрономическая обсерватория АКЦ ФИАН, Пущино
[2]Национальный исследовательский университет «Высшая школа экономики», Москва
[3]Институт космических исследований РАН, Москва
[4]Центральный научно-исследовательский институт машиностроения, Королёв
[5]ООО "Автоматизация Бизнеса", Москва
[6]Санкт-Петербургский государственный университет, Санкт-Петербург
sam@prao.ru , apozanen@iki.rssi.ru , rodin@prao.ru , dmitr22@list.ru , dumsky@prao.ru, eisaev@hse.ru , kaz.prao@bk.ru , lsv@prao.ru , oreshko@prao.ru , mt1710@yandex.ru , panther_gatchina@mail.ru

Аннотация

В Пущинской радиоастрономической обсерватории АКЦ ФИАН находится один из наиболее чувствительных радиотелескопов на частоте 110 МГц – БСА (Большая Сканирующая Антенна). Начиная с 2012 г. на БСА ФИАН стартовали непрерывные круглосуточные наблюдения на многолучевой диаграмме в полосе 109-112 МГц. Сейчас используются 96 лучей на склонениях от –8 до +42 градуса. Число частотных полос – от 6 до 32; постоянная времени 0,1 с и 0,0125 с. В режиме приема 32 полос с постоянной времени 0,0125 с за сутки регистрируется 87.5 Гбайт информации (в год 32 Тбайт). Эти данные дают большие возможности, как для краткосрочного, так и для долгосрочного мониторинга различных классов радиоисточников (в том числе радиотранзиентов различной природы), мониторинга ионосферы Земли, межпланетной и межзвездной плазмы, поиска и мониторинга различных классов радиоисточников. Для структурирования большого количества данных наблюдений и для систематизации их обработки была создана специализированная база данных (http://astro.prao.ru/cgi/out_img.cgi). В данной работе мы обсуждаем методы выделения импульсных явлений из нее и первые результаты такой обработки. При помощи базы данных выделено 83096 индивидуальных импульсных событий (на отрезке июль 2012 - октябрь 2013), которые могут соответствовать пульсарам, мерцающим источникам и быстрым радиотранзиентам. В результате создана однородная выборка импульсных явлений для последующего проведения статистического анализа обнаруженных событий.

Труды XVIII Международной конференции DAMDID/RCDL'2016 «Аналитика и управление данными в областях с интенсивным использованием данных», Ершово, 11-14 октября 2016

1 Введение

Пущинская радиоастрономическая обсерватория АКЦ ФИАН располагает одним из наиболее высокочувствительных радиотелескопов в мире - БСА (Большая Сканирующая Антенна или, в другой расшифровке – большая синфазная антенна). Радиотелескоп БСА ФИАН был создан в 1970 – 1974 г.г. БСА ФИАН является радиотелескопом меридианного типа и представляет собой эквидистантную фазированную антенную решетку, состоящую из 16384 волновых диполей расположенных на площадке 384x187 м (геометрическая площадь более 70 тыс. кв. м, эффективная – около 45 тыс.).

Радиотелескоп первоначально работал в диапазоне 101 - 104 МГц, но в 1996 году был перенастроен на диапазон 109 – 112 МГц (т.е. длина волны около 3 м). В этом диапазоне БСА является самым чувствительным телескопом в мире (и одним из наиболее чувствительных в мире на метровом диапазоне волн в целом). Системная эквивалентная плотность потока (SEFD) радиотелескопа равна 34 Ян [1] в зените при минимальной температуре фона, что примерно в 3 раза лучше, чем у радиотелескопа LOFAR [2] на частоте 110 МГц.

БСА ФИАН – это незаменимый инструмент для решения целого ряда задач в области исследования

пульсаров, изучения динамических процессов в околосолнечной и межпланетной плазме, анализа структуры компактных радиоисточников в метровом диапазоне волн. Заметим, что сейчас длинноволновая радиоастрономия (в диапазоне 10-300 МГц) переживает новую волну интереса. Для этого диапазона существует обширный класс научных задач (см., например, в [3]). Уже существующие инструменты (наиболее известный из них УТР-2 [4]) переживают второе рождение, строятся также новые радиотелескопы – кроме LOFAR нужно назвать ГУРТ [3] и LWA [5].

2 Наблюдения и аппаратура

2.1 Характеристики радиотелескопа БСА ФИАН

Важнейшей особенностью БСА является то, что он работает в режиме приема полной мощности. Это позволяет снимать на нем, помимо дискретных радиоисточников, фоновое излучение нашей Галактики и протяженные источники с характерным размером до 2-3 градусов. Вторая особенность – БСА изначально проектировался с возможностью формирования на нем многолучевой диаграммы. До 2007 г. диаграмма направленности БСА состояла из 16 лучей, затем была сформирована вторая, независимая 16-лучевая диаграмма. Наконец, в 2010-11 гг. была разработана и создана еще одна (уже третья) диаграммообразующая система из 128 лучей.

Основные технические параметры радиотелескопа БСА сейчас следующие:

- Рабочий диапазон частот 109 – 112 МГц;
- Эффективная площадь антенны (максимальная) – 47 000 м. кв.
- Системная температура шума (минимальная) – 560 К;
- Поляризация антенны – линейная (горизонтальная, вдоль направления восток-запад);
- Число одновременно формируемых лучей диаграммы – 128;
- Ширина луча диаграммы в Е-плоскости – 54 угл.мин.;
- Ширина луча диаграммы в Н-плоскости – 24 угл.мин.(зенит).

Новая система диаграммообразования формирует 128-лучевую диаграмму, плотно перекрывающую телесный угол в 40 кв. градусов. Но без цифровой многоканальной системой регистрации она недееспособна. Первые 48 лучей диаграммы были оснащены многоканальными приемниками 7 июля 2012 г. (старт круглосуточных наблюдений). Начиная с 1 апреля 2013 г. ведется запись файлов с наблюдательными данными уже с 96 лучей в секторе склонений $-8°<\delta<+42°$.

Характеристики приемно-регистрирующей системы (чувствительность – для зенита, при минимальной температуре фона):

- Число каналов (лучей) регистрации – 96;
- Полоса частот регистрации – 2,5 МГц с центральной частотой 110,25 МГц;
- Число частотных каналов в одном канале регистрации от 6 до 32;
- Интервал выборки сигнала в канале Δt=12,5 – 100 мс;
- Максимальная чувствительность (для всей полосы 2,5 МГц) радиотелескопа БСА при постоянной времени 100 мс равна 0,07 Ян, при 12,5 мс – 0,2 Ян;
- Чувствительность для одной спектральной полосы (415 или 78 кГц соответственно): 0,16 или 1,06 Ян для 100 и 12,5 мс соответственно.

Итак, начиная с 1 апреля 2013 г. на БСА ведется запись с 96 лучей по линии Юг в секторе склонений: $-8°<\delta<+42°$ (ранее, с 6.07.2012 - регистрация велась с 48 лучей, в секторах склонений $+8.7°<\delta<+12.6°$ и $25°<\delta<+42°$). Это составляет 5,08 стерадиана или 0,40 поверхности всей сферы. Многолучевая диаграмма сейчас одновременно покрывает 40 кв. градусов (1/1000 всей сферы).

При этом время прохождения источника через диаграмму БСА около 5 минут, т.е. вероятность попадания в произвольный момент времени источника из наблюдаемой полосы склонений в главные лепестки многолучевую диаграмму ~ 1/300.

Дополнительно особо отметим факт, что возможен захват сигнала от мощных источников также боковыми лепестками диаграммы в зоне наблюдений вплоть до ±40 градусов от главного лепестка. Действительно, форма диаграммы БСА описывается функцией

$$(\alpha,\delta) = [\sin x / x]^2 \cdot [\sin y / y]^2, \qquad (1)$$

где $x = (\pi D_1 / \lambda)(\alpha - \alpha_0)$, $y = (\pi D_2 \cos Z / \lambda)(\delta - \delta_0)$, λ и δ – экваториальные координаты, (λ_0,δ_0) – координаты, определяющие положение максимума диаграммы направленности радиотелескопа, D_1 и D_2 – размеры БСА ФИАН в направлениях с Востока на Запад и с Севера на Юг, соответственно.

Нетрудно найти из (1), что в боковых лепестках на указанном расстоянии от центрального направления регистрируется около 1/10000 от мощности сигнала в главном лепестке. Однако, ввиду того, что диаграмма антенной решетки имеет гораздо более сложный вид (с увеличенными боковыми лепестками), чем диаграмма идеального элемента, а также ввиду влияния погоды и прочих искажающих факторов, сигнал в далеких боковых лепестках увеличивается обычно на порядок и может достигать – 1/1000 от мощности сигнала в главном лепестке.

Как результат, в секторе около 1 стерадиана рядом с небесным меридианом БСА способен регистрировать на небе в любой момент времени источники порядка ~1 тыс. Ян (в том числе, особо отметим – источники транзиентного характера).

В случае прямого попадания в зону главного лепестка одного из лучей диаграммы, для критерия

обнаружения S/N≥10 у спорадического (транзиентного) сигнала на фоне обычных данных, получаем следующие оценки. Единичный надежно обнаружимый импульсный выброс данных на одной полосе 415 кГц (Δt=100 мс) будет ≥1,6 Ян, для полосы 78 кГц (Δt=12,5 мс – вполне удовлетворяет характерным масштабам FRB, особенно с учетом их возможного уширения ввиду межзвездного рассеяния на низких частотах) – ≥10,6 Ян. Отметим, что для проведенного недавно на LOFAR пульсарного обзора [6] авторы сделали аналогичную оценку с тем же критерием S/N=10 для поиска быстрых радиовсплесков FRB (см. далее), и она ≥107 Ян (для Δt=0,66 мс). То есть 32-частотные данные БСА с Δt=12,5 мс имеют хороший потенциал даже в сравнении с LOFAR, но при этом подобные обзоры мы делаем ежедневно. В данной работе, однако, для отработки методики мы пока работали с 6-частотными данными (Δt=100 мс).

2.2 Регистрация наблюдательных данных

Регистратор представляет собой промышленный компьютер с возможностью установки комплекта из 6 модулей для регистрации сигналов, идущих от разных лучей радиотелескопа БСА. Каждый модуль обрабатывает и регистрирует 8 сигналов (лучей). В итоге один регистратор записывает данные с 48 лучей радиотелескопа (работает 2 регистратора).

Сигнал от каждого луча разбивается на несколько частотных диапазонов. По умолчанию количество таких полос устанавливается равным 6 (максимально – 32 полосы) на общую полосу в 2,5 МГц. Кроме того, в выходной файл регистратора записывается также сигнал с общей полосы приема телескопа. При заданном количестве полос 6 (ширина каждой полосы 415 КГц), в выходном файле будет содержаться информация о 7 частотных диапазонах, при числе полос 32 (ширина каждой полосы 78 КГц) – соответственно записываются 33 числа для каждой временной точки с каждого луча. Обычно сигнал снимается с постоянной времени 100 мс, т.е. 10 раз в секунду (максимальная частота съема данных до 80 раз в секунду). Сами данные пишутся как 32-х битные числа с плавающей запятой. Итак, в стандартном режиме (6 полос по 415 КГц, с постоянной времени 100 мс) каждую секунду мы десять раз пишем массив четырехбайтовых чисел 48x7. В конце каждого часа накопившийся массив данных сбрасываются в файл (размером по 46 Мб с каждой из двух установок). В итоге, за сутки с 96 лучей диаграммы с апреля 2013 г. поступает 2.3 гигабайта научной информации (в год 848 Гбт).

В быстрой, «тяжелой» моде (32 полосы по 78 КГц, с постоянной времени 12,5 мс) соответственно в секунду пишется с одной установки 80x48x33x4 ≈ 507 Кб данных (в сутки с 2-х установок – 87.6 Гб, в год – 32 Тб). Именно данные в «тяжелой» моде особо ценны, поскольку могут использоваться для поиска радиотранзиентов самого различного вида – от пульсаров до FRB всплесков. Для поиска быстрых радиовсплесков (Fast Radio Burst (FRB) – открытые в 2007 [7] источники внегалактической природы, поскольку они имеют большие дисперсионные задержки DM>>100 pc•cm^{-3}) необходимы именно миллисекундные масштабы.

Поэтому в мае 2014 г. была проведена работа по перепрограммированию и настройке многоканального регистрирующего комплекса, после чего стала возможной регистрация одновременно двух режимов: и стандартного, и быстрого. В результате с июня 2014 ведется одновременная регистрация обеих мод. Соответственно, поток данных достиг почти 90 Гб в сутки и 33 Тб (терабайт) в год. На текущий момент (начало июля 2016) уже накоплено 4 года ежедневных обзоров для данных малого формата (6 полос 10 раз в секунду) и 2 года непрерывных наблюдений данных большого формата (32 полосы 80 раз в секунду). Всего сейчас уже хранится 68 Тб.

Этот поток данных является ключевым для мониторинга состояния земной ионосферы, мониторинга вспышек на Солнце (космической погоды), обнаружения тысяч мерцающих радиоисточников, мониторинга потоков сотен радиоисточников в нашей Галактике и за ее пределами, поиска новых радио объектов. В последнем типе задач наиболее интересны направления поиска быстрых радиотранзиентов (вспышек), т.н. FRB, на небе на масштабах миллисекунд. Подобные радиовсплески открыты в дециметровом диапазоне, но в метровом диапазоне их пока никто не находил. Мы полагали также возможным обнаружение новых радиопульсаров, и наши надежды уже оправдались [8,9]. Мы надеемся обнаружить в наших данных еще несколько десятков новых пульсаров.

Существуют и другие аспекты применения этих уникальных данных, в том числе для прикладных исследований. Однако для того, чтобы получить выход от этого огромного потока научных данных, его нужно структурировать и максимально эффективно обработать.

3 Обработка данных

3.1 База данных наблюдений

В феврале 2014 года сформирована специальная база данных (на основе PostgreSql) наблюдений на третьей многолучевой диаграмме БСА. Она снабжена средствами графического вывода данных. В базе данных были собраны наблюдения с 6 июля 2012 вплоть по 20-е октября 2013 г. для стандартной моды наблюдений [10]. Каждые 5 секунд этого периода рассчитывались 27 параметров для каждого луча, на их основе построено более 700 тыс. рисунков (в качестве примера см. рис. 1).

В феврале 2014 эти графические данные стали доступны на сайте ПРАО АКЦ ФИАН по адресу http://astro.prao.ru/, в декабре 2014 база данных была переведена в режим публичного on-line доступа.

Размещение сжатых данных в форматах реляционных баз данных позволяет использовать все доступные средства выделения, сортировки, сопоставления, фильтрации и начальной обработки данных при помощи механизма стандартных команд SQL. Это значительно упростило перекрестный и временной анализ данных, осреднение, выделение данных с нестандартным поведением и т.п.

Для всех 5-секундных точек данных (для каждого из 96 лучей, всего 8 млн. точек) получены: максимальные значения для данного луча $S_{max\_f\_n}$ и данной полосы (с номером n=1..6, при этом первая полоса на f=109.0 МГц, последующие через каждые 415 кГц, для n=7 – общая полоса 2,5 МГц) на этих 5 секундах; минимальные $S_{min\_f\_n}$, средние медианные $S_{med\_f\_n}$, дисперсии сигнала N_{f_n} и др.; всего 27 типов данных.

Отметим, что при нахождении средних значений $S_{med\_f\_n}$ – из каждых 50 последовательных значений отбрасывались 2 нижних значения и 5 верхних (с целью отбраковки "вылетающих" значений от коротких импульсов и предохранения от кратких сбоев аппаратуры); $S_{med\_f\_n}$ и N_{f_n} определялись по оставшимся 43 точкам. Отбраковка крайних значений не только делает оценку среднестатистических значений более достоверной, но и позволяет затем легко выделить данные, где есть краткие выбросы (см. далее 3.2).

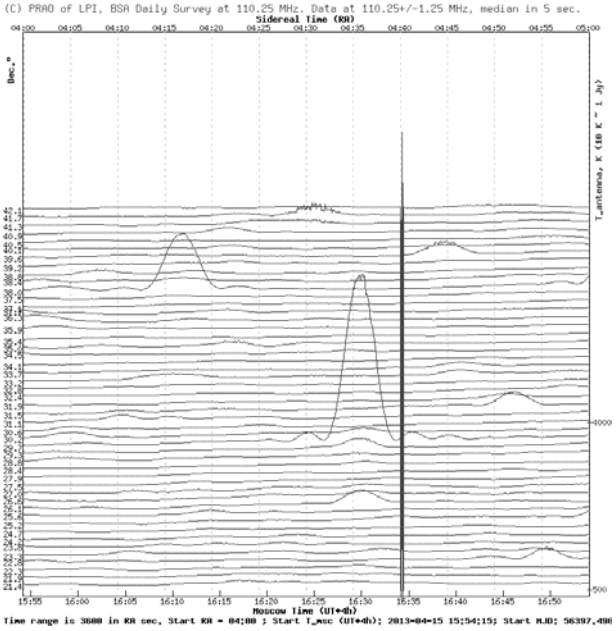


Рисунок 1 Часовой образец данных с многолучевой диаграммы БСА (на R.A.=4 h 40 m – калибровочные сигналы) для 15.04.2013 г. При прохождении радиоисточников через многолучевую диаграмму хорошо заметна форма луча диаграммы БСА (вместе с боковыми лепестками), явление мерцания у точечных источников и т.д.

База данных будет расширяться и перейдет в режим автоматического наполнения с непрерывной выкладкой новых графических данных на сайт.

3.2 Выделение импульсных данных

На основе использования базы данных создана методика выделения единичных импульсов на отдельных лучах диаграммы. Ее первый шаг состоял в подаче на базу данных логических команд на языке SQL, делающих выборку данных по комбинации логических условий:

– для полос n=1..6 искались выбросы на 5-ти секундных отрезках с критерием обнаружения $S_{f\_n} / N_{f\_n} \geq 5.0$, где $S_{f\_n} = S_{max\_f\_n} - S_{med\_f\_n}$;

– предыдущему условию должны удовлетворять не менее 3-х частотных полос из 6;

– затем проверялись далеко отстоящие лучи многолучевой диаграммы (далее, чем ±3 луча) – требовалось, чтобы они, напротив, не демонстрировали аналогичного $S_{f\_n} / N_{f\_n} \geq 5.0$. Данное условие отбраковывает большинство импульсных помех техногенного и природного характера (например, грозы), проявляющие себя обычно на всех лучах синхронно. (Хотя точность регистрации грозовых импульсов не позволяет использовать данные для триангуляции, т.е. совместного поиска местоположения каждого разряда, эти данные могут быть использованы для статистического анализа, например, совместно с космическими экспериментами для изучения грозовой активности.) Импульсы же небесного характера проявляют себя обычно лишь в 2-3 лучах многолучевой диаграммы БСА. Заметим, что в этом условии выборки может таиться и изъян методики: могут быть ошибочно отбракованы мощные импульсы на небе, зафиксированные далекими боковыми лепестками диаграммы направленности и проявившимися в более, чем на 3-х лучах.

Описанным образом при помощи базы данных было выделено 83096 отдельных импульсов для $+3.5° < \delta < +42°$ (июль 2012 – октябрь 2013). Каждый из них перечитывался из оригинального файла данных с целью углубленной обработки. Для двух полос с максимальными S/N находилось значения корреляции Пирсона для масштабов 1, 2, 5, 10 сек, массивы двигались друг относительно друга. Для уверенной корреляции при наличии заметных временных сдвигов между полосами (то есть DM>0) в хотя бы двух из четырех временных масштабов импульсу автоматически присваивалось признак импульса космического происхождения. Но, поскольку автоматический классификатор в 10-15% случаев промахивался, импульсы дополнительно проверялись и сортировались визуально (посредством специально разработанного веб-интерфейса). При классификации импульсов схожего характера, еще не описанных ранее, вводились их новые классы. Результаты сортировки сохранялись в базе данных в специальную таблицу.

Импульсы с дисперсионными задержками (космического происхождения) уже отсортированы. Пример такого импульса (отклик от одного из многочисленных импульсов пульсара PSR2305+3100) – представлен на рис. 2.

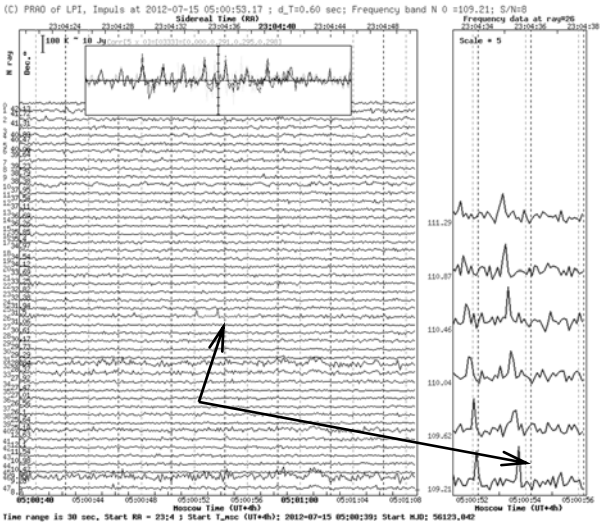


Рисунок 2 Пример одного из найденных импульсных событий: отклик от одного из импульсов (в центре рисунка) пульсара PSR2305+3000. Стрелками указан выделенный импульс: в левом, основном боксе рисунка – временная развертка по лучам диаграммы БСА, в правом боксе – развертка одного луча по 6 частотам. В правом боксе хорошо заметен сдвиг импульса вдоль частот (здесь DM=49,6 pc/cm^3). Верхний бокс рисунка показывает результат поиска корреляции Пирсона по 4-м временным масштабам: все они показали четкий сдвиг более низкочастотных данных относительно высокочастотной полосы. Данный пульсар достаточно ярок (~5 Ян) и заметна серия импульсов, следующих друг за другом с периодом P=1,58 сек. Всего для PSR2305+3000 на отрезке времени 6.07.2012 – 20.10.2013 гг. выделено 334 импульсных события.

Работа по классификации импульсов без ярко выраженных дисперсионных задержек (мерцания, аппаратурные и другие техногенные эффекты) будет закончена в ближайшее время. Пример одного такого импульса (пролет космической станции МКС через луч аэродромного маяка) показан на рис. 3.

4 Основные результаты

В результате работы сформирована однородная выборка индивидуальных импульсных событий, пригодная для проведения различных статистических исследований. Предварительно найдено, что:

А) 10-15% импульсов имеет техногенный характер различного рода: помехи, грозовые разряды, пролеты самолетов через диаграммы лучей, отражения аэродромных маяков от искусственных спутников Земли (пример – на рис. 3) и т.д.

Б) 15-20% – аппаратурные сбои данных (среди них могут содержаться реальные импульсы пока неопределенного происхождения – например, импульсы от пролета метеоров).

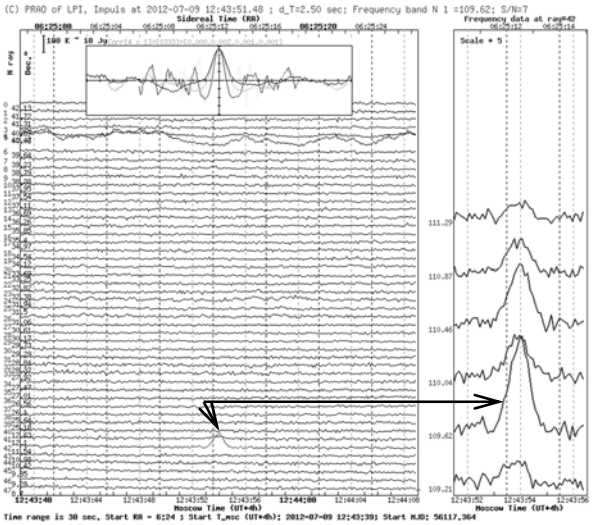


Рисунок 3 Пример одного из найденных импульсных событий: отклик от пролета спутника Земли (Международной Космической Станции, отражающая площадь около 400 м^2) через луч самолетного курсоглиссадного радиомаяка с одного из подмосковных аэродромов. Данные маяки работают на частотах около 110 МГц (попадают в диапазон частот БСА). Стрелками указан выделенный импульс (ширина около 1.5 сек): в левом, основном боксе рисунка – временная развертка по лучам диаграммы БСА, в правом боксе – развертка одного луча по 6 частотам. В правом боксе сдвига импульса вдоль частот – нет. Верхний бокс рисунка показывает результат поиска корреляции Пирсона по 4-м временным масштабам: все они показали, что корреляция максимальна при нулевом сдвиге от данных разной частоты наблюдений. Количество событий такого характера уточняется (ожидаем не менее нескольких десятков в нашей выборке). Вверху рисунка, на 6-м луче сверху – пример мерцаний точечного радиоисточника. Подобные сильные (одиночные) мерцания масштаба порядка секунды обычно также выделялись как независимые импульсные события.

В) около 25% импульсов имеет характер единичных мерцаний радиоисточников разного вида – от ионосферных до межзвездных (в том числе от боковых лепестков).

Г) около 40% импульсов имеет хорошо выраженный космический характер, показывая дисперсионное запаздывание импульсов относительно частоты (за счет прохождения в межзвездной среде).

Наиболее интересны результаты выделения импульсов космического происхождения. В частности, из примерно 340 пульсаров каталога ATNF [11], попадающий в полосу $+3.5°<\delta<+42°$ (наиболее пригодной для обработки) и с периодами более 0.3 секунды в нашей выборке найдены импульсы для 41 из них (еще 2 пульсара найдено из имеющих период 0.2-0.3 секунды). При этом число найденных импульсов для указанных пульсаров

колеблется от 1 (одного) до тысяч за все указанное время наблюдений (в среднем около года).

Для пульсаров с дисперсионным запаздыванием DM<15 импульсы зарегистрированы для каждого второго пульсара указанной выборки из ATNF, с дисперсионным запаздыванием DM<50 – для каждого третьего, для пульсаров 50<DM<100 – для каждого 25-го, для DM>100 в импульсах не найдено ни одного пульсара. Очевидно влияние уширения импульсов пульсара рассеянием на эффективность выделения импульсов из базы данных. Пути усовершенствования методики поиска импульсных явлений теперь довольно ясны (переход на данные с Δt=12.5 мс, увеличение времени анализа).

Найдено также несколько импульсов предположительно космического характера, по-видимому, никак не связанных с уже известными пульсарами.

Работа частично поддержана грантами РФФИ 14-07-00870а и 16-07-01028а.

Литература

[1] Орешко В.В. Радиотелескопы ПРАО – состояние и перспективы. http://www.prao.ru/conf/rrc2014/docs/22092014/06_Oreshko.pdf

[2] Van Haarlem M. P. et al: LOFAR: the low-frequency array // Astron. Astrophys. – 2013. – Vol. 556. – id. A2

[3] Коноваленко А.А. и др.: Астрономические исследования с помощью малоразмерных низкочастотных радиотелескопов нового поколения // Радиофизика и радиоастрономия. — 2016. — Т. 21, № 2. — С. 83-131.

[4] Брауде С. Я., Мень А. В., Содин Л.Г. Радиотелескоп декаметрового диапазона волн УТР-2 // Антенны. – М.: Связь, 1978. – Вып. 26. – С. 3–14.

[5] Taylor G. B. et al: First Light for the First Station

[6] of the Long Wavelength Array // J. Astron. Instrum. – 2012. – Vol. 1. – P. 1–56

[7] Coenen, Thijs et al: The LOFAR pilot surveys for pulsars and fast radio transients // Astronomy & Astrophysics, Volume 570, id.A60, 16 pp

[8] Lorimer, D. R. et al: A Bright Millisecond Radio Burst of Extragalactic Origin // Science. - 2007, Vol. 318, Is. 5851. — P. 777-780

[9] Тюльбашев С.А., Тюльбашев В.С. Открытие новых пульсаров на радиотелескопе БСА ФИАН. Астрономический циркуляр. 2015, N 1624, 4 pp.

[10] Родин А.Е., Самодуров В.А., Орешко В.В. Быстрый метод обнаружения периодических радиоисточников на БСА ФИАН. Астрономический циркуляр. 2015 , N 1629, 4 pp.

[11] Samodurov V. A., Rodin A.E., Kitaeva M. A., Isaev E., Dumsky D. V., Churakov D.D., Manzyuk M.O.: The daily 110 MHz sky survey (BSA FIAN): online database, science goals data processing by distributed computing. // Труды XVII международной конференции «Аналитика и управление данными в областях с интенсивным использованием данных» (“Data Analytics and Management in Data Intensive Domains” (DAMDID)). Обнинск: НИЯУ МИФИ, 2015. P. 127-128.

[12] G. Hobbs, R. N. Manchester and L. Toomey ATNF Pulsar Catalogue v1.54 : http://www.atnf.csiro.au/people/pulsar/psrcat/

The daily radio sky survey at 110 MHz: database and statistical analysis of transient phenomena in 2012-2013

V.A. Samodurov, A.S. Pozanenko, A.E. Rodin, D.D. Churakov, D.V. Dumsky, E.A. Isaev, A.N. Kazantsev, S.V. Logvinenko, V.V. Oreshko, M.O. Toropov, M.I. Volobueva

Radio Observatory of Lebedev Physical Institute in Pushchino has one of the most sensitive radio telescope at 110 MHz BSA (Big Scanning Antenna). Since 2012 BSA started multi-beams observations using 96 beams in declination from -8 up to +42 degrees in 6 and 32 frequency bands at 109-112 MHz and time sampling 0.1 s and 0.0125 s. The data stream in 32 bands and time sampling of 0.0125 s is producing 87.5 gigabytes per day. The data obtained can be used for short and long-term monitoring of various classes of radio sources (including radio transients), the Earth's ionosphere, interplanetary and interstellar plasma monitoring. A database is constructed to facilitate access to a large amount of observational data (see http://astro.prao.ru/cgi/out_img.cgi). We discuss algorithms of detection and identification of different classes of transients using the database. In particular we found 83096 events which could be associated with pulsars, scintillation sources and fast radio transients. These events are a homogenous sample suitable for statistical analysis.

Поиск иерархических звездных систем максимальной кратности

© Н.А.Скворцов © Л.А.Калиниченко
Институт проблем информатики ФИЦ ИУ РАН, Москва

© Д.А.Ковалева © О.Ю.Малков
Институт астрономии РАН, Москва

nskv@mail.ru leonidandk@gmail.com
dana@inasan.ru malkov@inasan.ru

Аннотация

В астрофизике кратных иерархических звездных систем существует противоречие между их максимальной наблюдаемой кратностью (6-7) и теоретическим ограничением на эту величину (до пятисот). Для поиска иерархических систем большой кратности проведен анализ современных каталогов как широких, так и тесных пар. Результатом работы является список объектов – кандидатов в звездные системы максимальной кратности, включающий тщательную кросс-идентификацию компонентов систем.

Работа проводилась при частичной поддержке РФФИ (гранты 16-07-01028, 16-07-01162, 14-07-00548).

1 Введение

Проблема кросс-идентификации небесных объектов возникает при работе над практически любыми задачами астрономии, и традиционно решается отдельно для каждого частного случая пересечения астрономических каталогов.

Для одиночных объектов эта проблема была осознана и решалась астрономическим сообществом с 80-ых годов прошлого века. Проблема кросс-идентификации двойных звезд заметно сложнее. Если для одиночной звезды это, как правило, только две координаты и блеск, то для двойной звезды учитываются координаты и блески главного и второстепенного компонентов, параметры их орбитального движения. Эта проблема обсуждалась астрономическим сообществом с конца 90-х годов прошлого века и была, в общих чертах, решена авторами статьи при создании Базы данных двойных звезд BDB (РФФИ 12-07-00528) [1, 2]. На сегодняшний день BDB – единственный ресурс астрономических данных, предоставляющий сведения о двойных звездах всех наблюдательных типов. Наконец, проблема кросс-идентификации объектов более высокой кратности разрабатывалась для ряда частных случаев. Решение этой проблемы в общем виде сталкивается с присутствием в системах одновременно объекты различных наблюдательных типов: изолированных (в эволюционном смысле) звёзд, переменных тесных затменных пар звезд, источников рентгеновского излучения, также указывающих на тесные взаимодействующие пары звезд, и ряда других. Соответственно, увеличивается число используемых для отождествления параметров объектов и особенностей их идентификации.

Одной из целей исследования очень кратных (very multiple) систем звёзд является поиск иерархических систем, подтверждающих теоретические обоснования возможности существования систем с определённым количеством уровней подчинённых пар звёзд. Эта проблема рассматривается в данной статье.

В разделе 2 описаны сущность теоретических ожиданий существования систем звёзд большой кратности и наблюдаемая картина реальных систем. Для исследования кратных систем в разделе 3 ставится проблема тщательного кросс-отождествления систем и их компонентов.

2 Теоретическая и наблюдаемая кратность звездных систем

2.1 Иерархические системы и теоретические ограничения на их кратность

Согласно современным представлениям тройная звездная система является динамически стабильной

Труды XVIII Международной конференции DAMDID/RCDL'2016 «Аналитика и управление данными в областях с интенсивным использованием данных», Ершово, 11-14 октября 2016

только в том случае, если она имеет иерархическую структуру, т.е. состоит из сравнительно тесной пары и удаленного компонента, составляющего с ней более широкую пару. При этом отношение периодов широкой и тесной пар должно превышать некое критическое значение, зависящее от эксцентриситета *e* внешней орбиты и равное 5 для случая круговой орбиты (для эксцентричных орбит это значение растет пропорционально $(1\text{-}e)^3$) [3]. Удаленный компонент также может представлять собой тесную пару звезд, и тогда данная конфигурация является примером иерархической четырехкратной системы.

Аналогично, наличие в такой звездной системе еще более удаленного компонента (третий уровень), орбитальный период которого не менее чем в 5 раз превосходит максимальный из уже имеющихся периодов, обуславливает появление иерархической системы более высокой кратности. Этот компонент также может оказаться двойным и т.д.

Следует заметить, что системы, не удовлетворяющие упомянутому выше ограничению на отношение орбитальных периодов, не являются гравитационно устойчивыми и динамически эволюционируют. Такая эволюция может включать сближения, выбросы звезд и заканчивается формированием **иерархической** системы исходной или меньшей кратности. Считается, что большинство одиночных и двойных звезд образовались как раз благодаря распаду неиерархических кратных систем [4].

Физический размер кратной иерархической системы ограничен сверху приливным влиянием гравитационного поля Галактики и случайными столкновениями с гигантскими молекулярными облаками. В [5] было показано, что число уровней иерархии не может превышать 8-9 (в зависимости от масс компонентов и орбитальным параметров пар). Следовательно, при максимально плотной «упаковке» кратность иерархической звездной системы может достигать значения 256-512 компонентов.

2.2 Наблюдаемая кратность иерархических систем

Одним из наиболее полных источников данных о кратных звездах является Каталог кратных систем MSC [6]. В каталог включены только иерархические (за редким исключением) и физические системы. Физические системы – это те, в которых гравитационная связь компонентов подтверждена их орбитальным движением или общим собственным движением (тангенциальным перемещением звезд на небесной сфере). Каталог MSC содержит около 1500 звездных систем кратностью от 3 до 7, причем из двух каталогизированных систем кратности 7 одна, по мнению автора, может являться молодым звездным скоплением (не обязанным демонстрировать иерархию членов).

Практическое отсутствие наблюдательных подтверждений существования систем кратности выше шести, которое демонстрирует содержимое каталога MSC, резко контрастирует с теоретическими оценками, приведенными в предыдущем разделе. Для ликвидации этого несоответствия необходимо привлечь дополнительные источники информации.

3 Отождествление кратных звёздных систем

3.1 Каталоги двойных и кратных систем

Таблица 1 Основные каталоги визуальных двойных и кратных систем.
C – количество компонентов,
P – количество пар,
S – количество систем,
M – кратность систем

	C, P, S	M
The Washington Double Star Catalog (WDS)	249280, 133966, 115314	2-44
Catalogue of Components of Double and Multiple Stars (CCDM)	105837, 56513, 49325	1-18
Tycho Double Star Catalogue (TDSC)	103259, 37978, 64869	1-11

Современные каталоги двойных и кратных звезд содержат системы гораздо более высокой кратности, чем семь. Это, прежде всего, WDS [7], CCDM [8], TDSC [9]. Сведения о них приведены в Таб. 1. Единицы, приведенные в последней колонке, указывают на (i) наличие в CCDM (некоторого количества) т.н. астрометрических двойных – систем, в которых второй компонент не наблюдается напрямую, но своим гравитационным влиянием модулирует собственное движение более яркого компонента, и (ii) на наличие в TDSC (изрядного количества) одиночных звезд, которые авторам каталога не удалось разрешить на подкомпоненты.

Нужно также отметить, что формально каталог WDS содержит несколько систем более высокой кратности, чем указано в Таб. 1, однако, они представляют собой либо набор звезд поля около центральной звезды (т.е., так называемые оптические пары, где компоненты располагаются на заметно отличающихся расстояниях, не связаны гравитационно и лишь проецируются в один участок небесной сферы), либо это – члены скопления, а не кратные системы.

При использовании информации, содержащейся в каталогах из таб. 1. необходимо учесть несколько обстоятельств.

Прежде всего, информация в каталогах WDS, CCDM, TDSC достаточно скудна, чтобы можно было делать окончательный вывод о физической связи конкретного компонента с системой (хотя, как будет

показано ниже, некоторые каталогизированные данные позволяют делать предварительные выводы на этот счет). Ни один из упомянутых выше каталогов не содержит данных обо всех известных звездах этого типа.

Каталоги также не свободны от ошибок: дупликации, включения одного и того же объекта (звезды) в разные системы, ошибок абсолютных и относительных координат, ошибок в значениях параметров, ошибок идентификации и других. Это можно проиллюстрировать на примере одной из систем, WDS 04078+6220 = CCDM 04078+6220 = TDSC 8749. Каталоги WDS, CCDM, TDSC содержат для нее сведения о 18, 16 (один из которых не включен в WDS) и 6 компонентах, соответственно, причем обозначения компонентов в системе различны (так, некий компонент имеет в этих трех каталогах обозначения O, S и D). Несколько звезд системы входят в другие каталоги: в одни – поодиночке, в другие – в паре. Детальный анализ этой системы вскрыл около 20 ошибок в семи различных каталогах и базах данных.

3.2 Алгоритмизация кросс-отождествления кратных систем

Проблема идентификации систем звёзд сводится к отождествлению многокомпонентных сущностей среди неоднородных данных из разных источников. Компоненты таких сущностей (систем звёзд) могут быть разных типов, отражая наблюдательные и астрофизические особенности звёздных объектов, входящих в состав систем, и соответственно характеризоваться разными наборами атрибутов (характеристик звёздных объектов), а также могут быть в свою очередь многокомпонентными в некоторых источниках данных.

Данные, доступные в наборе астрономических каталогов одиночных или кратных звёзд разных наблюдательных типов, анализируются для выявления одних и тех же компонентов звёздных систем, для их отождествления.

Идентифицированные кратные системы рассматриваются как сформированные на основании анализа данных связные графы, вершинами которых являются компоненты систем (либо звёздные объекты неразрешённые на сегодня на подкомпоненты), а дугами – рассматриваемые в каталогах пары компонентов от главного к второстепенному. Среди множества данных ряда астрономических каталогов необходимо корректно идентифицировать каждую вершину, каждую дугу и графы систем в целом. Очевидно, что ошибочное отождествление компонентов и пар в системах может повлечь за собой объединение нескольких систем в одну, причисление одиночных звёзд к системам и другие подобные ошибки.

Кросс-отождествление компонентов и пар между каталогами представляет определённую проблему: методика, описанная в [10], неплохо себя показавшая для систем кратности 2-3-4, зачастую пасовала перед системами большей кратности (т.е., в густонаселенных звездных полях) и требует проработки. Предлагаемый ниже подход к кросс-отождествления кратных систем основан на прежних методах, но призван исправить его недостатки, а также обеспечить анализ кратных систем с данными перспективных каталогов и потоковых ресурсов, пополняемых в режиме реального времени.

Реальные данные каталогов показывают, что при анализе данных для отождествления систем необходимо учитывать целый ряд проблем:

- различное форматирование данных в разных каталогах;
- различную семантику атрибутов в записях каталогов (например, координаты объекта в разных каталогах могут означать координаты фотоцентра пары или координаты более яркого из компонентов пары);
- ошибки ввода в каталогах (например, опечатки в идентификаторах идентифицированных звёзд в каталогах);
- отсутствующие значения в полях каталогов;
- изменчивые значения атрибутов (например, изменение блеска и координат между наблюдениями за счёт орбитального движения компонентов);
- неоднородность структуры комплексных объектов (например, компоненты неиерархической системы могут быть связаны в пары разными способами, а главными в паре – сочтены разные компоненты, если они имеют близкие характеристики);
- присутствие неструктурированных данных (указания в комментариях, полезные для идентификации объектов).

Таким образом, в решение задачи кросс-отождествления звёздных систем привлекается целый набор подходов к разрешению сущностей и слиянию данных. Используются разные наборы атрибутов и графовые структуры, на основе которых можно оценить идентичность систем и их компонентов. Отождествление может основываться не только на оценке параметров наблюдения и свойств объектов, но и учитывать идентификацию на основе уже идентифицированных объектов [11,12]. Всякую звёздную идентификацию, присутствующую в оригинальных каталогах в виде идентификаторов, ссылающихся на записи других каталогов, при возможности необходимо проверять с привлечением значений наблюдаемых параметров.

Методы должны быть применимы для решения задач отождествления кратных объектов в перспективных каталогах, а значит, ориентироваться не на особенности конкретных каталогов, как часто происходит при решении задач кросс-отождествления астрономических наблюдений, а на учёт обобщённых знаний предметной области об определённых типах астрономических объектов, об

особенностях разных методов их наблюдения, о влиянии характеристик оборудования на результаты наблюдений.

Работа по идентификации начинается с компонентов широких (визуальных) кратных систем. Разрешение многокомпонентных графовых сущностей, коими являются кратные звёзды, включает поиск дубликатов всех его составляющих частей во всём используемом наборе источников данных (каталогов и обзоров). Отождествляются друг с другом:

- вершины (компоненты систем) по атрибутам, а также на основании присутствия отождествлённых дуг и связи через дуги с другими вершинами;
- дуги (пары компонентов) по атрибутам, а также с учётом отождествлённых вершин;
- графы (системы звёзд) с учётом отождествлённых вершин и дуг.

Визуальные компоненты систем отождествляются, в первую очередь, методами, применяемыми при кросс-отождествлении одиночных звёзд. Для каждого компонента системы составляется множество его вероятных дубликатов во всех рассматриваемых каталогах (в том числе, и обзоров неба, не разделяющих объекты на одиночные или составные). Однозначная идентификация фиксируется при единственном элементе в множестве возможных идентификаций.

В множество попадают объекты на основании близости координат с учётом эпох наблюдения и собственного движения, а затем удаляются из множества те объекты, которые не соответствуют известным ограничениям предметной области, если необходимые для проверки данные об объектах присутствуют. Критериями могут являться: близость значений блеска или цвета (при известных фотометрических системах), собственного движения, тригонометрического параллакса, эволюционного статуса, спектральной классификации, и другие.

После обозначения множеств возможных идентификаций компонентов систем начинается фаза отождествления визуальных пар, которая должна внести новые критерии для устранения неоднозначностей идентификации. Для пар также составляются множества возможных идентификаций с парами компонентов из разных каталогов. В множество включаются все варианты перебора пар с учётом возможных идентификаций компонентов, составленных на предыдущем этапе. После этого, как и в случае с компонентами, к множествам возможных пар применяются известные ограничения предметной области и удаляются пары, не соответствующие критериям, если присутствуют данные для их проверки.

Положение вторичного компонента относительно главного в паре может различаться в различных каталогах из-за орбитального движения или из-за большой разницы собственных движений в случае оптической пары. Блески звезд могут заметно различаться в разных каталогах, если наблюдения проводились в разных фотометрических системах. Физическая переменность звезд также может привести к разным значениям блеска в разных каталогах.

Для каждой пары кандидатов на отождествление осуществляется сравнение значений позиционной и фотометрической информации. При этом для каждого атрибута (углового расстояния между компонентами, позиционного угла, блесков компонентов, разности блеска компонентов) по результатам статистического исследования каталогов определяется предельное возможное значение отклонения. Если разность значений атрибута не превышает предельного для этого атрибута значения, это служит критерием для отождествления пары.

Помимо этого, в некоторых случаях пару следует отождествлять не с парой другого каталога, а с компонентом. Одна и та же пара близких звезд, в зависимости от их блесков и углового расстояния, может быть каталогизирована при применении оборудования с разным угловым разрешением как один объект (с блеском яркого компонента или с интегральным блеском пары) или как два различимых объекта. Для определения таких ситуаций проводится определение фактического углового разрешения каталога, и в зависимости от него идентификация проводится с компонентом, либо с парой в целом.

Существует ряд методов, позволяющих выявлять оптические пары. Указанием на оптическую пару может служить заметная разница в значениях собственных движений компонентов и/или их годичных параллаксов (т.е., расстояний). Еще одним индикатором отсутствия гравитационной связи между компонентами пары, при наличии сравнительно длительного ряда наблюдений, служит линейное (а не орбитальное) относительное движение компонентов. Кроме того, известен статистический метод выявления вероятных оптических пар на основании плотности звездного поля в направлении галактических координат компонентов, блеска вторичного компонента и углового расстояния между компонентами (т.н. метод 1% фильтра [13]). Выявленные предположительно оптические пары отмечаются специальным флагом.

Вообще говоря, могут обнаруживаться звёзды из обзоров неба, которые подходят по параметрам, чтобы быть кандидатами в визуальные двойные, но не входят ни в один каталог двойных. Такие объекты отмечаются как кандидаты на вхождение в известные системы, либо как компоненты для составления новых систем. В множества возможных идентификаций пар добавляются пары с объектами, не входящими в каталоги двойных, но имеющими признаки двойных. Новые кандидаты пар с такими компонентами отмечаются особым флагом.

Составляются также правила, связанные с распространёнными ошибками или конфликтами в

каталогах. Например, разница калибровки блеска в фотометрических системах может быть предположена в случае, если блески объектов в разных каталогах отличаются на одну и ту же величину. Объекты, подходящие по критериям с учётом исправления ошибок, также включаются в множества возможных идентификаций с флагом типа возможной ошибки данных.

Однозначная идентификация пар возможна в случае, если после всех проверок в множестве для пары остаётся всего один кандидат на пару с другим каталогом. Такая пара фиксируется как идентифицированная. Пара удаляется из множества кандидатов на пары обоих компонентов. В результате, может появиться однозначная идентификация и для оставшихся пар. Также однозначная идентификация пары влечёт за собой и идентификацию её компонентов, так как участие в единственной возможной паре является существенным признаком идентификации. Идентифицированные компоненты удаляются из множеств возможных идентификаций других компонентов, в результате чего могут появиться новые однозначные идентификации других компонентов и пар.

На следующей стадии происходит подключение информации о более тесных системах, являющихся компонентами широких пар, исследованных выше. Эта информация включает данные о двойных/кратных системах следующих наблюдательных типов: интерферометрических, орбитальных, астрометрических, спектроскопических, затменных, рентгеновских, катаклизмических, двойных в радиопульсарах. Принципы отождествления базируются также на позиционной и фотометрической информации, но, вообще говоря, зависят от типа системы. Для каждого типа составляются свои ограничения предметной области, связанные со специфическими параметрами объектов. Также при отождествлении учитывается, что одни и те же пары могут фигурировать в разных каталогах как объекты разных наблюдательных типов.

Отождествление систем в целом осуществляется по наличию общих компонентов и пар. В одном участке неба могут находиться несколько систем, не связанных друг с другом, если их графы не связаны.

Наконец, на последнем этапе к полученным результатам кросс-отождествления компонентов и пар кратных систем добавляется информация об идентификации этих объектов в основных каталогах одиночных звезд (Bayer/Flamsteed, DM, HD, ОКПЗ, HIP; ссылки). Эти идентификаторы являются общепризнанными и широко используемыми. Однако вопрос о том, какому именно объекту соответствует тот или иной идентификатор, зачастую требует пристального рассмотрения. На данном этапе применяются правила, обнаруживающие разные типы ошибок идентификации. Например, предположение о перепутанных компонентах в паре может генерироваться, если в паре идентификаторы принадлежат разным компонентам в разных каталогах, а блеск компонентов в каталогах отличается на близкую по модулю величину, но с разным знаком.

Каждой системе, паре и компоненту назначается особый идентификатор, с которым связываются идентификаторы разных каталогов кратных и одиночных звёзд для формирования общей базы соответствий идентификаторов.

Не разрешённые автоматически множества компонентов и пар, а также элементы с установленными флагами новых объектов и разных типов ошибок рассматриваются экспертом.

4 Звездные системы кратностью 6+

4.1 Поиск физически связанных систем в каталогах визуальных двойных

Для окончательного решения проблемы кросс-отождествления очень кратных систем, а также для компиляции списка кандидатов в иерархические звездные системы максимальной кратности (и поиска значения этой максимальной кратности) нами была проделана работа по полуавтоматической идентификации систем кратности 6 и выше в каталогах из Таб. 1. Таких систем насчитывается 551, они включают в себя 5746 компонентов.

На первом этапе проводилось собственно кросс-отождествление компонентов системы в различных каталогах (кросс-отождествление самих систем было успешно осуществлено в [10], а их анализ приведен в [14]). При этом, как и ожидалось, был обнаружен ряд ошибок в оригинальных каталогах.

Далее, на основании значений каталогизированных параметров, выявлялись и помечались пары (члены систем), являющиеся оптическими. Указанием на оптическую пару может служить заметная разница в значениях собственных движений компонентов и/или их годичных параллаксов (т.е., расстояний). Еще одним индикатором отсутствия гравитационной связи между компонентами пары, при наличии сравнительно длительного ряда наблюдений, служит линейное (а не орбитальное) относительное движение компонентов. Для части систем эта информация включена в основную таблицу каталога WDS, для других должна извлекаться из текстовых примечаний к нему, на основании поиска и извлечения фрагментов текста по ключевым словам. Таким способом, с использованием критериев, связанных с движением компонентов, были обнаружены 1395 пар в 297 системах кратности 6+. Кроме того, статистический метод 1% фильтра позволяет заподозрить в оптической двойственности 2779 пар в 478 системах. Для 882 пар при этом действуют оба индикатора оптической двойственности.

Таким образом, число физически связанных компонентов в системах кратностью 6+ оказалось на

3292 ниже, чем общее количество компонентов, и составило 2454. Кратность 6+, после исключения из рассмотрения предположительно оптических компонентов, может быть приписана лишь 101 системе.

4.2 О неразрешенной двойственности компонентов кратных систем

Строго говоря, исследуемые системы могут иметь более высокую кратность, поскольку некий компонент системы (наблюдающийся как одиночная звезда) может оказаться, в свою очередь, двойной или кратной системой. Эта «скрытая», фотометрически неразрешенная двойственность может проявляться различными способами.

Так, если орбитальная плоскость такой тесной двойной развернута под достаточно большим углом к картинной плоскости, изменение лучевых (радиальных) скоростей компонентов вследствие орбитального движения проявляется в виде смещения спектральных линий компонентов в наблюдаемом спектре (эффект Доплера). Таких двойных (они называются спектроскопическими) на сегодняшний день известно около трех тысяч.

В случае же если наклон орбиты к картинной плоскости близок к 90 градусам, один из компонентов может в процессе орбитального движения проходить по диску второго (или затмевать его), что приводит к изменению интегрального блеска системы. Таких (т.н. затменных) систем известно, с разной степенью изученности, от семи до пятнадцати тысяч.

Наконец, самые тесные системы могут, вследствие эволюционного расширения одного из компонентов, перейти в стадию обмена веществом между компонентами. При этом «аккретор», если является очень компактным объектом (нейтронной звездой или черной дырой) не в состоянии аккрецировать сразу все вещество, поступающее от «донора». В системе образуется аккреционный диск, являющийся, вследствие градиента скорости вращающегося в нем вещества, источником рентгеновского излучения. Известно около четырехсот таких т.н. рентгеновских двойных.

В качестве примера можно привести упомянутую выше систему WDS 04078+6220 = CCDM 04078+6220 = TDSC 8749. Ее кратность увеличивается на четыре, если учесть, что один из ее компонентов представляет собой спектроскопическую двойную, а другой – четырехкратную систему, состоящую из двух еще более тесных пар: (i) спектроскопической и (ii) спектроскопической, наблюдаемой одновременно и как затменная.

Существует еще несколько менее представительных наблюдательных типов тесных двойных. Нужно отметить, что во всех случаях, перечисленных в этом разделе, наблюдатель имеет дело с одним источником света (т.е. компоненты не наблюдаются по отдельности).

Поиск тесных физических пар в кратных системах, наличие которых повышает уровень иерархии системы, проводился несколькими способами. Текстовые примечания к WDS (файл Notes) были разобраны для выделения информации о двойственном характере некоторых неразрешенных звезд, представленных в WDS как компоненты, но являющихся парой. Таким образом внутри систем высокой кратности были обнаружены 1 переменная двойная, 1 спектроскопическая двойная, и 33 тесных пар без указания наблюдательного типа. Кроме того, было проведено сопоставление с данными крупнейших каталогов спектральных двойных звезд (SB9, [15] – обнаружено 53 спектроскопических пары), переменных звезд (ОКПЗ, [16] – 19 затменных двойных) и орбитальных двойных (ORB6, [17] – 36 тесных пар, из которых 16 совпадают с найденными по Notes тесными парами без указания наблюдательного типа).

Итого были обнаружены 127 тесных пар, увеличивающих степень иерархии системы, в 92 системах. Дополнительные исследования должны быть проведены для того, чтобы определить в каждом из 35 случаев обнаружения в одной системе двух по-разному проявляющих себя фотометрически неразрешенных пар, разные ли это пары или одна и та же.

5 Заключение

Результатом работы является каталог отождествлений компонентов звездных систем высокой кратности, а также список систем, которые могут рассматриваться как иерархические системы наибольшей кратности. Этот последний список требует более тщательного анализа и дополнительных наблюдений.

Литература

[1] Kovaleva et al. 2015, Astronomy and Computing 11, 119

[2] Malkov et al. 2013, Astronomical and Astrophysical Transactions, 28, 235

[3] Tokovinin A., in Rev. Mex. Astron. Astrof. Conf. Ser., Ed. by C. Allen and C. Scarfe (Instituto de Astronomia, UNAM, Mexico) 21, 7, 2004.

[4] Larson R.B. The formation of binary stars: IAU Symp. 200. 93, 2001.

[5] Surdin V. ASP Conf. Ser. 228, 568, 2001.

[6] Tokovinin A., Astron. Astrophys. Suppl. Ser. 124, 75, 1997.

[7] Mason B.D., Wycoff G.L., Hartkopf W.I., Douglass G.G., Worley C.E. 2016, VizieR On-line Data Catalog: B/wds.

[8] Dommanget J., Nys O. 2002, VizieR On-line Data Catalog: I/274.

[9] Fabricius C., Hog E., Makarov V., Mason B., Wycoff G., Urban S. 2002, AAp, 384, 180.

[10] Isaeva A.A., Kovaleva D.A., Malkov O.Yu. 2015, Baltic Astronomy 24, 157.

[11] P. Christen. Data matching: concepts and techniques for record linkage, entity resolution, and duplicate detection. – Springer Science & Business Media, 2012. - ISBN: 978-3-642-31164-2. - XX+272 p.;

[12] I. Bhattacharya, L. Getoor. Entity resolution in graphs // Mining graph data. D. J. Cook, L. B. Holder (ed.) – John Wiley & Sons, 2006. – C. 311-332

[13] Poveda A., Allen C., Parrao L. 1982, ApJ, 258, 589

[14] Kovaleva D.A., Malkov O.Yu., Yungelson L.R., Chulkov D.A., Gebrehiwot Y.M. 2015, Baltic Astronomy 24, 367

[15] Pourbaix, D., Tokovinin, A.A, Batten, A.H., et al. 2014, VizieR On-line Data Catalog: B/sb9

[16] Samus, N.N., Durlevich, O.V., et al. 2013, VizieR On-line Data Catalog: B/gcvs

[17] ORB6: Mason and Hartkopf 2007, IAUS 240, 575

Search for hierarchical stellar systems of maximal multiplicity

Nikolay A. Skvortsov, Leonid A. Kalinichenko, Dana A. Kovaleva, Oleg Y. Malkov

According to theoretical considerations, multiplicity of hierarchical stellar systems can reach, depending on masses and orbital parameters, several hundreds. On the other hand, observational data confirm an existence of at most septuple systems. We study very multiple (6+) stellar systems from modern catalogues of visual double and multiple stars, trying to find candidates to hierarchical systems among them. Some of their components were found to be binary/multiple themselves that increases system's degree of multiplicity. Also, to collect all available information on those systems it was first necessary to make a thorough and accurate cross-identification of their components.

Стендовые и демо презентации

Posters and Demo

Когнитивная визуализация графических образов информационных объектов в технологиях интеллектуального анализа данных

А.В. Мышев А.В. Дудин
Национальный исследовательский ядерный университет (МИФИ)
Обнинский институт атомной энергетики (ИАТЭ), Обнинск, Россия,
mishev@iate.obninsk.ru

Аннотация

Рассматриваются методы когнитивной визуализации графических образов информационных объектов (ИО) как новой парадигмы для разработки информационных и компьютерных технологий интеллектуального анализа больших объемов данных нового поколения типа Brainware. Сегменты данных любого типа (тексты, видео, звук и др.) в каналах хранения и передачи информации любой сетевой вычислительной системы и телекоммуникационной системы представляются как бинарные множества, на которых заданы логические структуры. Содержательно–смысловая посылка логических схем методов состоит в следующем: 1) возможность перехода от традиционных форм представления вербально–символьной дискретной бинарной информации и ее восприятия к образному, используя информационную насыщенность и когнитивность графических образов (ГО); 2) математический и логический аппарат пространственных преобразований ГО позволяет выявлять неизвестные закономерности и новые знания, сокрытые в них; 3) проявление метаморфичности синергии зрительного восприятия и когнитивной аналитики больших потоков данных.

1 Пространственные преобразования ГО – основа когнитивной визуализации

Основная посылка метода пространственных преобразований ГО состоит в том, что исходную дискретную размытую функцию P(S,T) преобразуем в другую функцию P(S,k) такого же типа и далее исследуем ее ГО на основе метода компьютерной топографии [1,2]. Кратко поясним в виде определений суть компьютерной топографии.

Компьютерная топография ГО в системе базовых понятий и определений предметной и проблемной областях когнитивной графики является такой сущностью, посредством которой излагается аналитика и логика методологии разработки и реализации технологий визуализация для исследования информационных объектов (ИО). Обозначенные ИО аналогичны таким объектам, как стохастические поля в физике или поверхности в картографии [3,4,5]. В области практической реализации компьютерной топографии ГО в виде компьютерных технологий когнитивной графики основная посылка этой сущности состоит в том, что она в определенной мере, используя схемы и подходы компьютерной томографии [6,7,8], а также идеи исследования случайных функций по проекциям [9,10], отражает новую парадигму разработки средств научной визуализации на основе сетевых моделей когнитивных технологий синтеза и анализа ГО. Обобщенное определение понятия компьютерная топография ГО можно сформулировать следующим образом.

Компьютерная топография ГО – это метод анализа трехмерных ГО размытых дискретных функций, геометрия значений которой представляет собой перспективную решетку в трехмерном пространстве, на основе проекций ее значений либо на координатные плоскости, либо на любую гиперплоскость трехмерного пространства. Эти проекции представляют собой в памяти среды когнитивных технологий научной визуализации ГО область структурированных данных (СД), которые являются информационными картами (ИК) картографических моделей (КМ) проекций ГО. На экране монитора ИК отражаются в виде разноцветной карты, которую можно интерпретировать как картографический образ (КО). Тогда множество подмножеств клеток КО одинакового цвета образует морфометрию “ландшафтной” структуры КО. КМ обладают рядом фундаментальных свойств, благодаря

Труды XVIII Международной конференции DAMDID/RCDL’2016 «Аналитика и управление данными в областях с интенсивным использованием данных», Ершово, 11-14 октября 2016

которым исследователь получает представление о конфигурации и пространственной структуре (внешней и внутренней) в пределах масштаба и шкалы. К этим свойствам можно отнести следующие, а именно: 1) содержательное соответствие, которое определяется уровнем полноты и достоверности исходной информации; 2) абстрактность – наличие математической модели; 3) избирательность – возможность декомпозиции КМ с целью ее анализа; 4) синтетичность КМ обеспечивает целостность изображения; 5) метричность КМ обеспечивается математическим законом проекции, точностью синтеза и воспроизведения карты; 6) однозначность определяется отношением изоморфизма и гомоморфизма; 7) непрерывность изображения; 8) наглядность, как фактор образного восприятия КМ; 9) обзорность, как фактор выявления скрытых закономерностей. Следует также отметить факторы зрительного восприятия КО, которое ведется на трех уровнях: 1) восприятие частей КО; 2) восприятие всего изображения КО; 3) анализ изображения КО.

Содержательно–смысловая составляющая назначения функции P(S,k) состоит в том, чтобы описать и отразить в графической форме относительные частоты или вероятности появления значений функции P(S,T) относительно значений размытой функции S(T). Параметрическая переменная k дискретной функции S(k) является атрибутом–идентификатором, посредством которого определяется номер подинтервала, множество которых покрывает ось PO системы координат PSOT. А дискретная функция S(k) отражает зависимость значений размытой функции S(T) от значений переменной k, т.е. отражает эволюционную неопределенность значений размытой функции S(T).

Процедура определения масштаба разбиения оси Ok на подинтервалы реализуется в виде следующей алгоритмической схемы. Значениями дискретной размытой функции P(S,T) на пространственной перспективной решетки являются относительные частоты вероятностей, определяемых на σ – алгебре событий оси OS.

Нумерация подинтервалов оси Ok организована таким образом, что интервал с номером 1 включает малые значения вероятности, включая минимальное, а подинтервал с наибольшим номером – высокие значения вероятности, включая максимальное.

2 Вместо заключения

Методы и процедуры технологий когнитивной визуализации ГО на основе пространственных преобразований и их компьютерной топографии можно рассматривать как новую парадигму в технологиях научной визуализации и интеллектуального анализа информационных объектов в DID на когнитивном уровне их зрительного восприятия. Многочисленные практические реализации рассматриваемой парадигмы при решении различных задач [4] позволили выявить уникальные особенности исследуемых ИО [3,4 и др.].

Литература

[1] Пилюгин В., Маликова Е., Пасько А., Аджиев В. Научная визуализация как метод анализа научных данных / Научная визуализация. 2012. Т.4. №4. С. 56–70.

[2] Мышев А.В. Теория компьютерного восприятия и технологии взаимодействия вычислительного интеллекта с виртуальной средой моделирования // “Кибернетика и высокие технологии XXI века”: труды седьмой международной научно-технической конференции (16–18 мая 2006 г., Воронеж, Россия): в 2 т., т.2. – Воронеж: изд. ВГУ, 2006, с. 497–509.

[3] Мышев А.В. Компьютерная топография графических образов в технологиях интеллектуального анализа и распознавания информационных объектов // “Системный анализ и информационные технологии”: труды четвертой международной конференции (17–23 августа 2011г., Абазково, Россия): в 2 т., т.2. – Челябинск: изд. ЧГУ, 2011, С. 184–190.

[4] Мышев А.В. Модели активной памяти в технологиях виртуализации каналов передачи и хранения информации / Программные продукты и системы 2010. №1. С. 54-58.

[5] Берлянт А.М. Образ пространства: карта и информация. М.: Мысль, 1986, 240 с.

[6] Мышев А. В. Информационная модель нейросети в технологиях вычислительного интеллекта и формах реализации компьютинга / Информационные технологии, 2012, №1, с.62–70.

[7] Хермен Г. Восстановление изображений по проекциям. Основы реконструктивной томографии. М: Мир, 1983, 352 с.

[8] Наттерер Ф. Математические аспекты компьютерной томографии. М: Мир, 1990, 288 с.

[9] Ушаков В.Г., Ушаков Н.Г. Восстановление вероятностных характеристик многомерных случайных функций по проекциям. Вестник Московского университета. Серия 15. Вычислительная математика и кибернетика, 2001, №4, с. 32–39.

[10] Шестаков О.В. Влияние погрешностей в проекционных данных на алгоритм восстановления распределения случайной функции из распределений ее проекций. Вестник Московского университета. Серия 15. Вычислительная математика и кибернетика, 2002, №2, с. 35–40.

Cognitive Visualization of Graphic Patterns of Information Objects in Technologies of Data Mining

Alexey V. Myshev, Andrey V. Dudin

The methods of cognitive visualization of graphic patterns of information objects (IO) as a new paradigm for the development of information and computer technologies data mining of large volumes odata such as a new generation of Brainware is described. Segments of any type of data (text, video, sound, etc.) in the channels of the storage and transfer of information of any network computer system and the telecommunications system are represented as binary sets on which the logical structure is defined. The content and semantic posting logic method is as follows: 1) the ability to transition from traditional forms of verbal and symbolic representations of discrete binary information and its perception to the figurative representation, using information and cognitive richness of graphic patterns (GP); 2) mathematical and logical apparatus of GP spatial transformations can detect unknown patterns and new knowledge, hidden in them; 3) the manifestation synergy of visual perception and cognition analytics for large data streams.

Web tool for heuristic table structure recognition in untagged PDF documents

© Alexey Shigarov © Andrey Mikhailov © Andrey Altaev
Matrosov Institute for System Dynamics and Control Theory of SB RAS,
Irkutsk, Russia
shigarov@icc.ru mikhailov@icc.ru altaev@icc.ru

Abstract

PDF is one of the most popular document formats. Many PDF documents are untagged. They have no tags for identifying the logical reading order, paragraphs, figures, and tables. One of the challenges with these documents is the table extraction, including detection and structure recognition. We propose a new web tool for table structure recognition in untagged PDF documents. It is driven by a set of customizable conditions and ad-hoc heuristics for recovering table rows, columns, and cells. The experimental results based on the recognized competition dataset show high recall and precision for our tool.

1 Introduction

Today, PDF is one of the most popular document formats in the web. Many PDF documents are not images, but remain untagged. They have no tags for identifying the logical reading order, paragraphs, figures and tables. One of the important challenges with these documents is how to extract tables from them. *Table extraction* typically consists of two main steps: *table detection*, i.e. recovering the bounding box of a table in a document, and *table structure recognition*, i.e. recovering its rows, columns, and cells.

Many existing methods and tools for extracting tables from documents traditionally deal with only images or plain-text as source. They can be applied to PDF documents through converting PDF to these formats. However, this leads to loss of useful information. Table extraction from PDF directly can provide better results in comparison with these tools. There are only a few of tools intended to PDF table extraction. Some of them are discussed in the recent surveys [1, 4].

Our web tool exploits features of table presentations in untagged PDF documents. We use customizable conditions and ad-hoc heuristics for recovering table cells from text chunks and rulings. Most of them such as horizontal and vertical distances, fonts, and rulings are well known and used in the existing tools. Additionally, we have examined the possibility of applying the order of the appearance of text chunks in PDF files for table structure recognition. Usually, when a table printed in a PDF document originally was an object (e.g. a table in a Word-document) then (i) one printing instruction forms a part or a whole of textual content of only one physical cell; (ii) printing instructions forming a text inside each physical cell appear in the PDF file in the order that coincides with the human reading order of this text. We notice that it is true for many PDF generators. This feature can be especially useful in case of multi-row cells in table heads without rulings. It is important to note, that this has been done for the first time.

2 Methodology

We present the process of table structure recognition as three consecutive steps: (i) preprocessing, i.e. generating and preparing text chunks and rulings from a source document (Fig. 1); (ii) text block recovering, i.e. combining text chunks into text blocks (Fig. 2); and (iii) cell recovering, i.e. dividing table space into rows, columns, and cells via text blocks (Fig. 3).

2.1 Preprocessing

We operate two kind of objects: text chunks and rulings. A text chunk consists of a text, bounding box, and font characteristics. They present all machine-readable text in PDF documents. A ruling is a graphic line. To generate them we interpret PDF instructions for printing text and vector graphics via iText[1] library.

Initially each text chunk corresponds to one instruction of text printing. The same text can be presented in PDF by different printing instructions, depending on the used PDF generator, as shown in Fig. 1, *a*–*c*. At first, we split all text chunks (Fig. 1, *a*) into one-character chunks (Fig. 1, *b*) and merge them into word chunks with removing space characters and reindexing the order of their appearance (Fig. 1, *c*). We exclude each text chunk when it contains only one character marking itemized lists (e.g. bullet, square). Often, this kind of text chunks is visually detached from the rest of text chunks by long spaces. This lead to improperly recovered columns. Thus, eliminating them, we try to prevent some errors.

Visual rulings can be originally presented by printing instructions for graphic lines and rectangles. We merge all segments of one visual line (Fig. 1, *d*) into one ruling

Proceedings of the XVIII International Conference «Data Analytics and Management in Data Intensive Domains» (DAMDID/RCDL'2016), Ershovo, Russia, October 11 - 14, 2016

[1] http://sourceforge.net/projects/itext

(Fig. 1, *e*) and split each rectangle (Fig. 1, *f*) into four rulings corresponding to its boundaries (Fig. 1, *g*).

Fig. 1. Preprocessing of text chunks (*a*–*c*) and rulings (*e*–*g*).

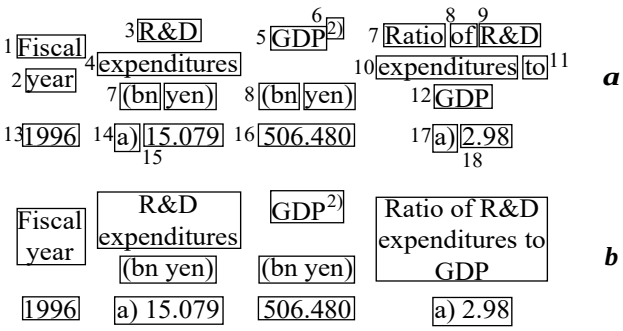


Fig. 2. Text chunks and the order of their appearance (*a*), and text blocks constructed from them (*b*).

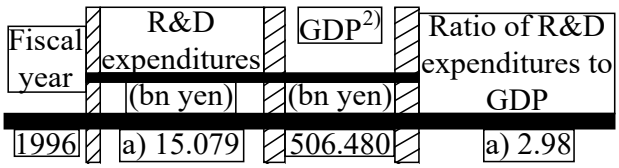


Fig. 3. Detecting vertical (indicated by the hatch pattern) and horizontal (indicated by the black solid pattern) whitespace gaps between text blocks to divide table space into cells.

2.2 Text block recovering

2.3

We suppose that each *text block* is a textual content of one cell, and each non-empty cell contains only one block. On this stage, we try to recover non-empty cells without their arrangement in rows and columns. All text chunks are combined into blocks (Fig. 2). One chunk can be included only in one text block. Text chunks are handled in pairs.

Our tool is driven by configurable conditions and ad-hoc heuristics. The most of them specify making decisions on combining text chunks into text blocks in the stage of text block recovering. In the web implementation, we define that two text chunks can be combined into one block when they satisfy the following conditions in case of horizontal (vertical) combining: (i) the horizontal (vertical) distance between the chunks is less than a space width of the left chunk (a height of the upper chunk); (ii) there is an intersection of their vertical (horizontal) projections.

We also use the following ad-hoc heristics: two text chunks combining into one text block have to satisfy the following conditions: (i) they are adjacent in the order of their appearance in the source PDF file; (ii) there are no rulings in the rectangle between them; (iii) they have identical font family, size, bold and italic attribute.

2.3 Cell recovering

Cell recovering is based on the whitespace analysis. We use the algorithm [5] to recover horizontal and vertical gaps between text blocks. Each whitespace gap corresponds to an implicit ruling. Thus, we try to recover all rulings, which separate cells in a table.

3 Experimental results

To evaluate our tool we use the existing competition dataset, "ICDAR 2013 Table Competition" [3], in its part regarding to the table structure recognition sub-competition. The dataset contains 156 tables in PDF documents collected from the web[2]. We also use the existing methodology for evaluating algorithms for the table structure recognition in PDF documents proposed in the paper [2]. The evaluation was performed automatically using Nurminen's Python scripts[3] for comparing ground-truth and result files that implement this methodology with slight modifications. The experimental results show high recall 0.9121 and precision 0.9180 for our tool.

4 Conclusions

The main applications of our tool are in the field of data accessibility, information extraction, and unstructured data integration. The further work is in progress on expanding the set of ad-hoc heuristics. We believe the involvement of the additional features such as text alignment, superscript, and subscript will allow improving our tool. In the future, our system also can be extended for supporting PDF table detection. The presented web tool is available at http://cells.icc.ru/pdfte.

This work was financially supported by the Russian Foundation for Basic Research (grants 15-37-20042, 14-07-00166) and Council for Grants of the President of Russian Federation (grant NSh-8081.2016.9).

References

[1] Coüasnon B., Lemaitre A. Handbook of Document Image Processing and Recognition, chapter Recognition of Tables and Forms, pp. 647–677, 2014.

[2] Göbel M., Hassan T., Oro E., Orsi G. A methodology for evaluating algorithms for table understanding in PDF documents. In Proc. of the 2012 ACM Symposium on Document Engineering, pp. 45–48, 2012.

[3] Göbel M., Hassan T., Oro E., Orsi G. ICDAR 2013 table competition. In Proc. of the 12th Int. Conf. on Document Analysis and Recognition, pp. 1449–1453, 2013.

[2] http://www.tamirhassan.com/dataset

[3] http://tamirhassan.com/competition/dataset-tools.html

[4] Khusro S., Latif A., Ullah I. On methods and tools of table detection, extraction and annotation in PDF documents. J. Inf. Sci., 41(1): 41–57, 2015.

[5] Shigarov A., Fedorov R. Simple algorithm page layout analysis. Pattern Recognition and Image Analysis, 21(2): 324–327, 2011.

Разработка алгоритмов автоматизированного определения жанрового типа и стилистической окраски текстов на русском языке

© В.Б.Барахнин © О.Ю.Кожемякина © И.С.Пастушков

Институт вычислительных технологий СО РАН,
Новосибирский государственный университет
bar@ict.nsc.ru olgakozhemyakina@mail.ru pas2shkov.ilya@gmail.com

Аннотация

В работе излагается алгоритм автоматизированного определения жанрового типа и семантических характеристик текстов на русском языке, основанный на создании словарей стилистически дифференцированных лексем.
Работа выполнена при частичной поддержке Президиума СО РАН и президентской программы «Ведущие научные школы РФ» (грант 7214.2016.9).

1 Введение

В задачах автоматизированного анализа текстов на естественном языке возникает проблема определения их жанрового типа и семантических характеристик. С этой проблемой исследователь может столкнуться в широком спектре ситуаций: от задач автоматизации комплексного анализа поэтических текстов, для которых жанровый тип и семантические характеристика являются важными атрибутами, используемыми при определении влияния низших уровней стиха на высшие (см., например, [1]), до отслеживания сообщений в социальных сетях с целью выявления террористических угроз, определения маркетинговых предпочтений покупателей и т.п.

.2 Построение классификаторов

Первым этапом решения этой проблемы является разработка соответствующих классификаторов. Для текстов на русском языке принято восходящее к трудам М.В.Ломоносова [2] деление текстов (прежде всего, художественных) на относящиеся к высокому, нейтральному и низкому стилям. Исторически каждый из них характеризуется 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. В свою очередь, в классической теории жанр произведения строго диктует выбор того или иного стиля. Так, применительно к стихотворным текстам эта зависимость подробно рассмотрена в работе [3]. Для других типов текстов нами разработан совместный («двумерный») классификатор жанровых типов и стилистической окраски текстов. Такой классификатор создается впервые (по крайней мере, для текстов на русском языке).

3 Создание словарей стилистически дифференцированных лексем

Следующий этап исследований – создание словаря лексем с выявлением их стилистической дифференциации.

Большое внимание вопросам стилистической дифференциации слов уделено в монографии [4]. Приведены списки слов «разговорных», со «сниженной» стилевой характеристикой и с «повышенной» стилевой характеристикой. Разумеется, 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, в течение шести десятилетий, прошедших с момента публикации монографии, некоторые включенные в списки слова сменили стилистическую окраску. Поэтому для соотнесения слова с тем или иным стилем нами используется предложенный в той же монографии 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.

Труды XVIII Международной конференции DAMDID/RCDL'2016 «Аналитика и управление данными в областях с интенсивным использованием данных», Ершово, 11-14 октября 2016

Ситуация осложняется 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 абсолютно надежным критерием отнесения текста к определенному жанровому или стилистическому типу, поскольку большинство лексем имеет полисемантическую окраску. Поэтому в перспективе для повышения точности алгоритма планируется выявление контекста вхождения той или иной лексемы, то есть описание семантических полей, соотносимых с различными жанровыми и стилистическими типами текстов, практически реализуемая также в виде словаря.

4 Выбор алгоритма классификации текстов

Следующая стоявшая перед нами задача – выбор метода классификации текстов на основе вхождения в них лексем, соответствующих тем или иным классам. Среди работ в этой области можно отметить статью [5], в которой проанализированы методы классификации на основе анализа распределения лексических дескрипторов естественного языка, а также представлен метод классификации текстовых документов на основе характеристики тематической значимости. Однако этот метод наиболее эффективен для работы с короткими документами. В итоге мы остановили свой выбор на алгоритме AdaBoost, впервые предложенном в работе [6], который на основан принципе усиления простых классификаторов: сначала пространство входных векторов отображается в пространство значений простых классификаторов, составляется их линейная комбинация, а решение принимается в зависимости от ее знака. Иными словами, пространство значений простых классификаторов разделяется гиперплоскостью, а принятие решения зависит от того, по какую сторону гиперплоскости находится отображение вектора признаков. Алгоритм характеризуется высокой скоростью работы, простотой реализации и высокой эффективностью распознавания.

В настоящее время ведутся вычислительные эксперименты с корпусом разножанровых и разностилевых текстов с целью демонстрации практической работоспособности рассматриваемого алгоритма.

5 Заключение

Таким образом, на основе использования перечисленных выше словарей стилистически дифференцированных лексем разрабатывается и реализуется в виде прототипа веб-приложения алгоритм автоматизированного определения жанрового типа и семантических характеристик текстов на русском языке.

Литература

[1] В. Б. Барахнин, О. Ю. Кожемякина. Об автоматизации комплексного анализа русского поэтического текста. Труды Четырнадцатой Всероссийской научной конференции «Электронные библиотеки: перспективные методы и технологии, электронные коллекции» (RCDL’2012), с. 213–217, Переславль-Залесский, 15-18 октября 2012 г.

[2] М. В. Ломоносов Предисловие о пользе книг церковных в российском языке. В: Ломоносов М. В. Полн. собр. соч. Т. 7. М.; Л.: Изд-во АН СССР, 1952. С. 585-592.

[3] Д. М. Магомедова. Филологический анализ лирического стихотворения. М.:Издательский центр «Академия», 2004.

[4] О. С. Ахманова Очерки по общей и русской лексикологии. М.: Учпедгиз, 1957.

[5] Э. Мбайкоджи, А. А. Драль, И. В. Соченков. Метод автоматической классификации коротких текстовых сообщений. Информационные технологии и вычислительные системы. 2012. № 3. С. 93–102.

[6] Y. Freund, R. E. Schapire. A decision-theoretic generalization of jn-line learning and an application to boosting. Journal of computer and system sciences. 1997. V. 55. P. 119-139.

The development of algorithms of the automated determination of the type of genre and stylistic coloring of Russian texts

V.B.Barakhnin, O.Yu.Kozhemyakina, I.S.Pastushkov

In this paper we describe the algorithm of automated definition of the genre type and semantic characteristics of texts in Russian based on the creation of dictionaries of stylistically differentiated lexemes.

Contextual Search as a Technique for Extracting Knowledge on the Internet

© Victor Telnov

National Research Nuclear University MEPhI,
Obninsk
telnov@bk.ru

Abstract

The subject of the demonstration is the pilot project devoted to creation and usage in educational activities of universities a retrieval system, which provide to students, teachers and professors a means of contextual search of educational resources on the Internet.

Proposed and implemented an adaptive technology of systematization and clustering of the found network content on the basis of the Pareto dominance relation under the joint consideration of many aspects (relevance, pertinence, etc.). Fuzzy text comparison is performed by means of the Levenshtein metric.

1 Introduction

Why the additional search agent is necessary if there are public search engines (Google, Yandex, Yahoo, etc.)? We will note several characteristic features of the majority of popular search engines which are well known to most of users:

1. The found documents are ranked by a search engine in accordance with its internal algorithm, which is not always in the interests of a particular user.
2. It is not easy to users to manage a search context, to specify and to guide the searching.
3. References to «promoted» commercial sites usually have a higher rating in comparison with other search results.

The above circumstances makes the public search engines less and less adequate tool for extracting knowledge on the Internet for educational purposes.

The discussed project is realized as part of more general educational portal «Department Online» [1].

Access to the agent «Contextual search» is possible via the homepage of the educational portal on the network address http://ksst.obninsk.ru or on a direct reference http://ksst.obninsk.ru/semantic.

2 Related work and novelty

Externally, the agent «Contextual search» looks like a metasearch engine because it uses the resources of many search engines to increase the probability of finding the desired information and to ensure completeness of search results.

Proceedings of the XVIII International Conference «Data Analytics and Management in Data Intensive Domains» (DAMDID/RCDL'2016), Ershovo, Russia, October 11 - 14, 2016

If not to consider early (created to the middle of the 2000th years) rather simple metasearch engines, then among the modern online services, are designed to extract knowledge from a global network, it is necessary to mention the search portal **AskNet.ru**, which is essentially a self-learning analytical question-answer system. **AskNet.ru** makes the linguistic analysis of texts (morphology, syntax, semantics) with application of ontologies, knowledge bases and inference methods.

Another intelligent search engine, **Нигма.рф** has its own database and provides additional indexing of online content, allowing the user to refine a search query, group together and filter search results for specified topics.

Deserves special mention the search–analytical system named **Exactus Expert**, aimed to support of scientific and educational activities, which is designed to study specific subject areas and analyze of the quality of scientific publications. **Exactus Expert** particularity is that the content is not extracted from the global network in the online mode, but is extracted from the own collection of documents with volume of about 2 million units, and user interface of the system is located in the so-called Deep Web.

The considered «Contextual search» supplements a line of intellectual search engines, offering adaptive technology of sorting and a clustering of the found network content on the basis of the Pareto dominance relation [2] at the joint accounting of many aspects, including relevance and a pertinence of data. Fuzzy comparison of texts is carried out with use of distance of Levenstein [3]. The search context in this agent is understood not as a traditional formal triplet of the form {objects, attributes, incidence}, but as some union of the taxonomies of educational portal as well as arbitrary fragments of texts, network documents and sets of keywords, which can be managed by the user during the implementation of the search, see Fig. 1.

3 System overview

Global search of documents and retrieval of specialized resources is initially performed by regular search engines (Google, Yahoo, Yandex, Mail.ru), the interaction with which occurs asynchronously through a standard API. The built-in query languages for regular search engines are used in full. The received content is some kind of raw material for further processing in the agent «Contextual search». Specifically, for each snippet is calculated its relevance, pertinence and a number of other indicators (aspects), which are used in order to categorize the search results.

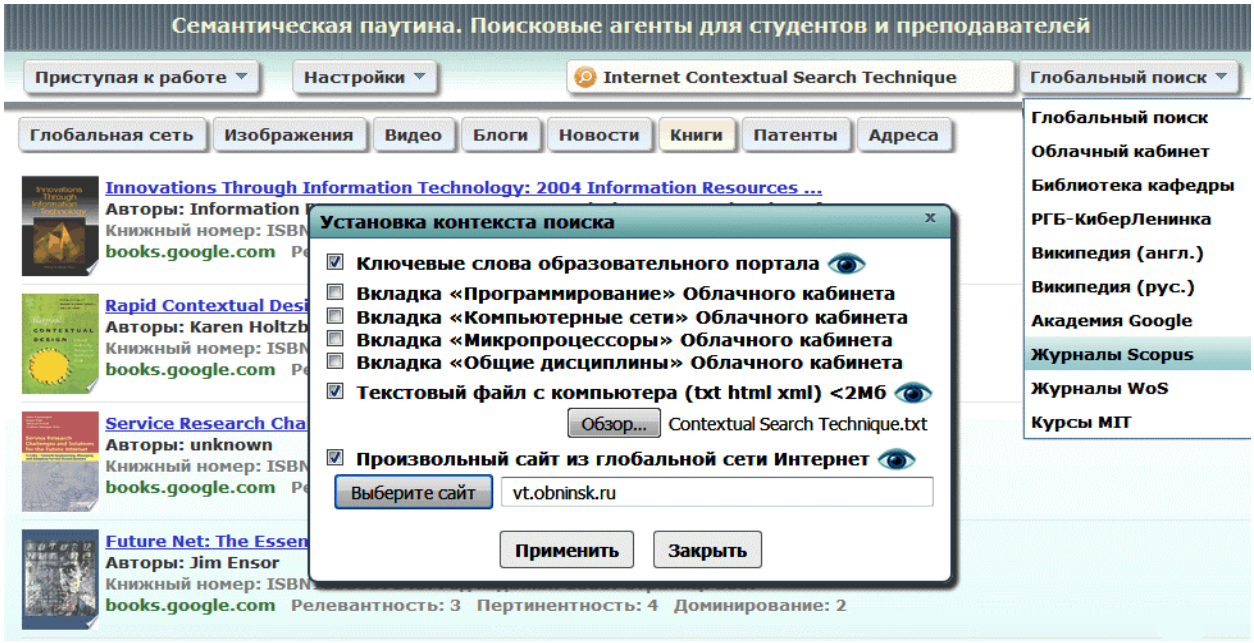


Figure 1 Setting a context during implementation of the search

Here relevance of a snippet is understood as a compliance measure between the text of a snippet and a text of a search query. The pertinence of a snippet is understood as a compliance measure between the text of a snippet and search context, as he has been defined above. These and other measures are calculated by a fuzzy comparison of the relevant texts, it is allowed a subtle customizing of the prices (weights) of individual Levenshtein operations, which are initially (by default) all equal to 1.

Worthy of special mention such aspect as Pareto dominance index, which is calculated on the basis of Pareto dominance relation and ensures the joint consideration of many other aspects that characterize the adequacy of snippets. The snippets with the maximum value of Pareto dominance index form the so-called Pareto set. The Pareto set includes the snippets, having the best combination of all the considered aspects, including the relevance and pertinence. Groups of snippets with an identical Pareto dominance index form clusters, which in the final output are located in decreasing order of this index. There are more ways to streamline and organize the found content, including any combinations of the aspects that characterize the adequacy of the snippets.

4 Demonstration overview

During the demonstration, results of work of the agent «Contextual search» are compared with similar search results, which received by popular search engines (Google, Yahoo, Yandex, Mail.ru).

Configuring the system is available both in the main panel of crawler, and also through the menu item «Settings». The completeness and adequacy of the search results is determined by the options «Search Engines», «Search Context», «Sort results».

Fine customizing of the mechanisms of Pareto dominance relation and the computation of the Levenshtein metric is demonstrated through management of the options «Pareto dominance index», «Tokens and Metrics».

The system saves all individual settings made by each user, so repeated setting of options isn't required.

Acknowledgements

The work was supported by the NBO «Vladimir Potanin Charity Fund», project № ГК160001360.

References

[1] V.P. Telnov, A.V. Myshev. «Department online»: a cloud computing in higher education. Programmnye produkty i sistemy, 2014, No. 4, pp. 166-172. http://www.swsys.ru/index.php?page=article&id=3903

[2] Ralph Keeney, Howard Raiffa: Decisions with multiple objectives: Preferences and value tradeoff. J. Wiley, New York (1976)

[3] Vladimir I. Levenshtein. Binary codes capable of correcting deletions, insertions, and reversals. Soviet Physics Doklady, Vol. 10, No. 8. (1966), pp. 707-710

Определение потенциала продаж розничных магазинов с использованием информации о других магазинах и гео-данных

© Н. И. Клемашев © И. В. Комаров © Б. А. Позин

ЗАО «ЕС-лизинг»
Москва
nklemashev@ec-leasing.ru ikomarov@ec-leasing.ru bpozin@ec-leasing.ru

Аннотация

В статье предлагается подход к построению модели расчета потенциала продаж магазинов. Расчеты модели могут использоваться как при разработке стратегии, так и для решения тактических задач улучшения отдачи от работы торговых представителей.

Основой предлагаемого подхода является алгоритм из двух шагов. Первый шаг – построение кластеров на основе общедоступных гео-данных и данных о магазинах. Гео-данные собираются о характеристиках магазина, которые влияют на покупательский спрос. Соответственно, кластеры объединяют магазины, которые максимально похожи друг на друга по спросу. На втором шаге выделяются лидеры кластеров по информации о продажах магазинов, входящих в кластер. Потенциал рассчитывается по средним продажам лидеров.

1 Построение кластеров магазинов

Часто можно услышать, что для успеха розничного бизнеса главное – это место, место, место. Действительно, проходимость, количество людей, живущих поблизости, наличие мест притяжения населения, количество конкурирующих магазинов поблизости [1, 2] и т.д. – являются определяющими факторами при открытии новых магазинов. Если говорить о предметах повседневного спроса (FMCG), то удобство расположения магазина, при условии незначительной разницы в ценах с конкурентами, является решающим фактором при выборе магазина для покупки товара.

1.1 Создание признаков

С развитием гео-сервисов, таких как OpenStreetMap, Гугл Карты и Яндекс Карты, а также других специализированных порталов с гео-информацией о фирмах (2ГИС, Yell.ru), данные о транспорте, местах притяжения населения, и даже количестве домов, их длине и этажности, становятся доступными для массовой обработки алгоритмами. Гео-данные собираются специализированными поисковыми роботами [3].

Для данных с гео-координатами (транспортные остановки, места притяжения населения) можно рассчитать расстояние до ближайшей точки и общее количество точек в определенном радиусе, например, 1 км, от магазина. Созданные таким образом переменные используются для построения признаков, которые определяют принадлежность к кластеру.

Помимо информации о потенциальном спросе, вторым важным набором данных является информация о магазинах. Магазины можно классифицировать по формату (супермаркет, традиционный магазин с прилавком, дискаунтер и т. п.) и торговой площади.

1.2 Алгоритм

Самым распространенным алгоритмом кластеризации, пожалуй, является алгоритм K-средних. Среди достоинств алгоритма можно выделить его простоту, масштабируемость и интерпретируемость. К недостаткам можно отнести предположения, касающиеся вида кластеров (выпуклость), необходимость задания количества кластеров, отсутствие единого объективного показателя качества кластеризации (что является проблемой всех алгоритмов кластеризации), чувствительность к выбросам (так как вычисляются средние).

В настоящей работе применяется алгоритм кластеризации K-средних. Массив данных предварительно разбивается на группы по значениям категориальных признаков (тип магазина и размер города, в котором находится магазин). Алгоритм K-средние применяется к каждой группе отдельно.

2 Расчёт потенциала

2.1 Определение потенциала

Под потенциалом понимаются возможности магазина по достижимым объемам продаж, определяемые в предложенной модели как средние

Труды XVIII Международной конференции DAMDID/RCDL'2016 «Аналитика и управление данными в областях с интенсивным использованием данных», Ершово, 11-14 октября 2016

продажи лидеров кластера, к которому принадлежит магазин. Потенциал можно измерять в рублях, в категориях товара или количестве единиц категории товара.

Абсолютный потенциальный прирост продаж рассчитывается как разность между потенциальными и фактическими продажами. При отрицательных значениях, абсолютный потенциальный прирост продаж принимается равным нулю.

Помимо абсолютного потенциала интерес представляет относительный потенциал, определяемый как отношение абсолютного потенциального прироста продаж к фактическим продажам.

2.2 Лидеры кластеров

Для кластеров возникает задача определения лидеров. Лидеры определяются как магазины с наибольшими продажами, при этом продажи у этих магазинов распределены по наибольшему количеству категорий. Для определения количества лидеров используем процент от всего количества магазинов, например, 10%.

3 Развитие модели

Описанный подход строится на определённых допущениях, которые требуют проверки.

Для проверки допущений и улучшения алгоритма собираются фактические данные о продажах за период, следующий за отчетным. На основании данных о продажах оценивается работа торговых представителей, которые руководствуются расчетами модели в своей работе.

При росте эффективности работы торговых представителей с использованием рекомендаций модели, т.е. при росте продаж, при прочих равных условиях, можно говорить об эффективности представленной модели.

Литература

[1] Esri, Inc. Huff Model. https://www.arcgis.com/home/item.html?id=f4769668fc3f486a992955ce55caca18

[2] J. Mitríková, A. Šenková, S. Antoliková. Application of the Huff Model of Shopping Probability in the Selected Stores in Prešov, the Slovak Republic. Geographica Pannonica, Volume 19, Issue 3, 110-121 (September 2015).

[3] Новые методы работы с большими данными: победные стратегии управления в бизнес-аналитике: Научно-практический сборник / А. В. Шмид и др., М.: ПАЛЬМИР, 2016. – 528с.

Estimating Potential of a Retail Shop Using Data of Other Shops and Geo-Data

Nikolay I. Klemashev, Ivan V. Komarov, Boris A. Pozin

The paper describes an approach for estimation of potential sales for a retail shop using data from other shops and readily available geographic data. Distances to the nearest shop, the number of shops in vicinity, geo-data of locations attracting visitors, geo-data of transportation points as well as characteristics of shops are used to construct clusters with similar potential sales. The quality of clusters is assessed using a custom k-means metric. Sales data is used to determine cluster leaders, average sales of which are used as a benchmark to assess potential of other shops in the cluster.

The approach is subject to field verification and improvement, using feedback from sales representatives

Диссертационный семинар

PhD Workshop

Unevenly Spaced Spatio-Temporal Time Series Analysis in Context of Volcanoes Eruptions

Grigory Trifonov

Moscow State University, Moscow

trifonov.grigory@gmail.com

Abstract

Paper presents a simplistic approach towards the detection and dynamics analysis of volcanic eruptions represented as unevenly spaced spatio-temporal time series of satellite retrieved hot spots. The paper discusses isolation and interpolation of hot spots data produced by Nightfire algorithm for the purposes of short-term volcanic activity ARIMA based forecasting. The case study for Chirpoi Snow volcano is presented.

The work is supported by RFBR (grants 14-07-00548, 16-07-01028, 15-29-06045).

1 Introduction

Volcanic eruptions are one of well-known types of natural hazard. Some volcanoes are better studied than the others, this mostly depends on their location, and those located on uninhabited islands may erupt without people in field observing them, while the others may erupt nearby cities and attract thousands of people. There are different kinds of volcanic activity such as: gas emissions, ash plumes, lava flows and domes, etc. These events may accompany one another. Some volcanoes demonstrate regular patterns in their activity others not. Events listed may be tracked by instrumental networks consisting of different instruments like seismographs, regular and thermal cameras, gas analyzers and others. Such networks sometimes are impossible to deploy due to various reasons. However it is possible to utilize satellites based instruments; such instruments do not require any special onsite installation and provide coverage for the entire planet.

1.1 Volcanoes satellite observations

As of 2016 there is no single satellite developed specifically for volcanology purposes. Fortunately there are a lot of meteorological satellites suitable for volcanology purposes. There are basically three kinds of satellite data which has a widespread use for volcanology purposes, these are:

1. visual data — which is basically just an image;
2. infrared data — most of volcanic events are accompanied with heat emission, hence they are visible in infrared spectra;
3. radar readings — lava flows and domes may be observed as they change the landscape.

This paper describes an approach towards infrared data analysis.

1.2 Current problem state

There are at least two global monitoring systems of volcanic activity based on hot spots detection in satellite data, namely: MODVOLC [4], MIROVA [3]. The oldest one (MODVOLC) has been in service for more than a decade by now [8]. Both systems are using MODIS sensor data as an input, which is provided by two NASA's polar-orbiters Terra and Aqua, launched in 1999 and 2002 respectively. Each service is using hot spots detection algorithm of its own. Essentially any such algorithm is based on the Plank curves fitting with some variations and modifications, book “Thermal Remote Sensing of Active Volcanoes” by Andrew Harris [2] provides a good introduction on how does it work. Both services provide real time and historical hot spots attributed to volcanoes and alerts based on some radiant heat threshold value. Neither of services does any additional hot spot analysis.

1.3 Suggested approach

This paper suggests an alternative approach towards volcanic activity monitoring and dynamics analysis. Instead of just tracking hot spots it is suggested to analyze hot spots as time series for each volcano. Such an approach will let to better adjust alert thresholds as well as to look up behavioral patterns and anomalies for all the existing volcanoes rather than just to detect potential eruptions.

2 Data

This study is using hot spots detected by a Nightfire algorithm [1]. Hot spots come in a form of a separate csv file for each 24 hours; each csv file contains around 100 fields out of which the most important and suitable for analysis are: pixel longitude and latitude, radiant heat, estimated black body temperature, estimated black body area, cloud conditions, satellite angles, detection quality flags and original image geometry. Due to its nature the data used have some specificity, which are worth consideration.

2.1 Satellite imaging

Nightfire is using VIIRS sensor nighttime data to detect hot spots. VIIRS sensor is basically a next improved

Proceedings of the XVIII International Conference «Data Analytics and Management in Data Intensive Domains» (DAMDID/RCDL’2016), Ershovo, Russia, October 11 - 14, 2016

generation of MODIS sensor, as of April 2016 there is one polar-orbiting satellite carrying VIIRS sensor operating (Suomi-NPP launched in 2011). Satellite continuously goes from one Earth pole to another following sun terminator. Every full circle satellite is crossing equator once in a day time and once in a night time. Satellite is constantly reading pixels in a line orthogonal to its path, each pixel size at Nadir (directly under the satellite) is 742x776 m, this size is non linearly growing towards the edge of scan, scan width is 3040 km, the satellite does 14 full revolutions each 24 hours, which essentially means that every point at equator is seen at least once every day and once every night. However for higher latitudes due to a considerable swath overlap locations are seen several times each day and each night.

2.2 Hot spot sources

There are several major sources of hot spots, anthropogenic ones, such as: gas flares over oil fields, high temperature manufacturing, thermal power stations as well as natural ones of which the most significant are forest fires and volcanic activity. Coincidentally hot spots of volcanic and forest fire origins have closely overlapping temperature ranges.

Nightfire has been originally geared towards the detection and tracking of gas flares over oil fields. Gas flares are smaller than VIIRS pixel (they either lay inside one pixel or between several neighboring pixels); hence Nightfire performs subpixel analysis, to estimate heat source size and energy. In the case of volcanic events it is no rare occasion to have dozens of hot-spots per image, which are all part of the same lava field for instance [7]. Thus it is crucial to somehow group such hot spots as the ones attributed to a single event.

2.3 Missing data

Another problem apart from separation of hot spots of volcanic and non-volcanic origins is a high number of missing or incomplete readings. There are several reasons for volcanic hot spots to become corrupted or even missing:

- clouds – may cover the area of interest, this leads to either missing hot spots or hot spots with a lower radiant heat than it should be;
- volcanic gasses and ash – these act more or less the same as clouds;
- volcano being too far off nadir – cauldron like volcanoes with lava lake inside crater may produce hot spots only at angles close to nadir, since otherwise satellite's sensor can't see lava lake;
- polar day – since Nightfire works with nighttime data only there will be no hot spots detected for the entire duration of polar day.

2.4 Data specifics summary

To summarize, there are three key points about data involved:

1. there are hot spots readings spanning since the early 2012 available for each volcano;
2. readings are unevenly spaced in time with a distance varying between 2-24 hours (discounting polar day/night cases);
3. each day of the observations comes in a form of a set of hot-spots with varying coordinates, scan time, radiant heat and temperature.

The three key points mentioned allowed to classify data as a set of unevenly spaced spatio-temporal time series.

3 Analysis

To observe volcano as a process it is first necessary to bisect the data related to the volcano in question. The simplest method to attribute hot spot to a volcano is to check how far it is from it. Basically the largest thermal anomalies, which can be seen from space, are lava flows and these typically would not reach further than 20 km from the volcano summit, hence distance based hot spots filtering leaves out most of the hot spots of non-volcanic origins. This simplistic approach does not protect from false attribution of forest fires to a volcanic activity, however as it has been already mentioned forest fires temperature range closely overlaps with that of volcanic events, hence such discrimination is a subject of future research.

3.1 Interpolating data

To interpolate time series for a specific volcano it is first necessary to somehow regularize the data presented within a single reading. As was already mentioned each reading may consist of several hot spots of varying radiant heat, temperature, area, cloud conditions all of them with different coordinates and satellite angles. There could be no more than a single hot spot for each pixel of the original satellite image. Knowing satellite altitude, nadir and azimuth angles as well as how does internally VIIRS sensor work [6] it is possible to calculate the area of the original pixel for each hot spot. A pixel area may be calculated with the following formula:

$$\alpha = asin\left(\frac{R \times sin(\pi - \beta)}{H + R}\right)$$

$$px_s = \Lambda R \frac{px_{ns}}{H}\left(\frac{cos(\alpha)}{\sqrt{\left(\frac{R}{R+H}\right)^2 - sin(\alpha)^2}}\right)$$

$$px_t = \Lambda \frac{px_{nt}}{H}\left(cos(\alpha) - \sqrt{\left(\frac{R}{R+H}\right)^2 - sin(\alpha)^2}\right)$$

Where R – earth radius, β – satellite zenith angle, H – satellite height, px_{ns} – pixel along scan size at nadir (776m), px_t – pixel along track size at nadir (742m), Λ – aggregation group which is $\frac{1}{3}$ if $\alpha \geq 44.68º$, $\frac{2}{3}$ if $\alpha \geq 31.589º$ and 1 otherwise.

This formula produces an upper boundary area estimate, while hot spot area calculated by Nightfire

itself could be used as a lower boundary estimate. Next step would be to aggregate both boundaries within any given reading:

$$A_{RU} = \sum_{p} A_{PU}$$
$$A_{RL} = \sum_{p} A_{PL}$$
$$RH_R = \sum_{p} RH_p$$
$$T_R = \sum_{p} A_p \times \frac{T_p}{A_R}$$

Where A_{RU} – reading active area estimated upper bound, A_{PU} – hot spot area upper bound, A_{RL} - reading active area estimated lower bound, A_{PL} - hot spot area lower bound, RH_R – reading's aggregate radiant heat, RH_p – hot spot radiant heat, T_R – estimated reading temperature, T_p – estimated pixel temperature.

The majority of the time series analysis theory is geared towards the analysis of regularly spaced time series; hence there are essentially two ways to analyze an unevenly spaced time series. The first one would be to interpolate an unevenly spaced time series and proceed with analysis of a regularly spaced time series. The second method would be to analyze unevenly spaced time series without performing such a transformation. For this paper the first approach as the simpler one has been chosen. Thus the resulting unevenly-spaced time series are linearly interpolated into their evenly-spaced form.

3.2 Analyzing interpolated time series

It comes as no surprise that in the majority of cases it would be impossible to forecast eruption dynamics due to the lack of data and high irregularity of the processes involved. However some volcanoes may show some regularity in their activity, which may signalize an applicability of classic time series forecasting techniques. One such popular technique ARIMA model widely used in econometric will be used in the case study to attempt to forecast volcanic activity of a selected volcano.

ARIMA (autoregressive integrated moving average) is widely used in econometrics. The model uses an initial differencing step to reduce non-stationarity of data combined with both autoregressive and moving-average models. Model is usually denoted as $ARIMA(p, d, q)$ where p – the order of autoregressive model, d – is the degree of differencing, q – is the order of moving-average model. $ARIMA(p, d, q)$ model is given by:

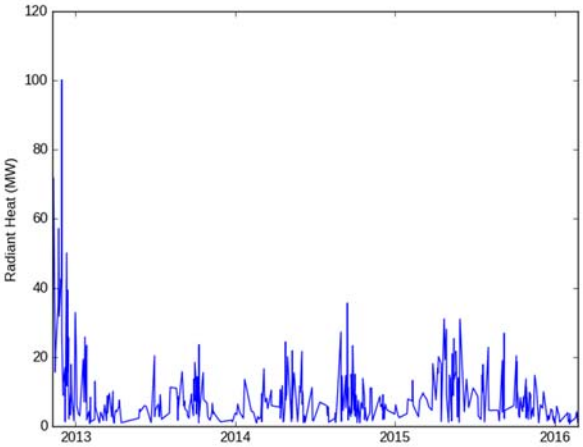
$$\left(1 - \sum_{i=1}^{p} \phi_i L^i\right)(1-L)^d X_t = \left(1 + \sum_{i=1}^{q} \theta_i L^i\right)\epsilon_t$$

3.3 Forecasting Chirpoi Snow activity

Snow is a stratovolcano located on an uninhabited volcanic island Chirpoi. The island is located in the Sea of Okhotsk between Simushir and Urup in the Kuril island chain. Snow has been continuously erupting for a since April 2012, its eruption seems to follow a pattern, where periods of activity are interleaved with short periods of low to no activity. There is no other dedicated observing equipment but the only unmanned seismic station, which may be disabled for weeks due to battery discharge, located on the island. Nightfire has the full data for this eruption and there are no other potential sources of hot spots, which could have added additional noise. Factors mentioned make Snow a good candidate for a case study.

This case study attempts to answer the question: “Is it possible to forecast eruption power output, at least for some volcanoes?” Power and radiant heat are interchangeable in this context. Power poses the most interest since there is a direct relation between radiant heat observed and the volume of lava being erupted.

Radiant heat readings Figure 1 exhibit non stationary behavior, however after differencing step Figure 2 situation gets better. For a difference time series mean is 0 standard deviation is 11.38. In fact early weeks of eruption impact standard deviation. Trimming first 6 weeks off time series brings standard deviation to about 8 and this value is valid for almost every segment of time series. Thus after differencing step process shows strong signs of being near stationary, which means that d parameter of ARIMA model will be 1 in this case. For p and q start parameters 3 will be taken.



Fugure 1 Chirpoi Snow Radiant Heat (MW) December 2012 - February 2016

To perform ARIMA forecast Pythons StastModels [5] package has been used. Model has been fit on 100 readings and attempted to predict the next 24 hours, next the closest to 101 reading hour prediction has been used to cross validate. One hundred retrains in the middle of the data set were made. Even though in most cases ARIMA managed to perform forecast sometimes it was impossible due to random shocks, which disrupted data strong enough to make data set non stationary, hence

ARMA step is not applicable.

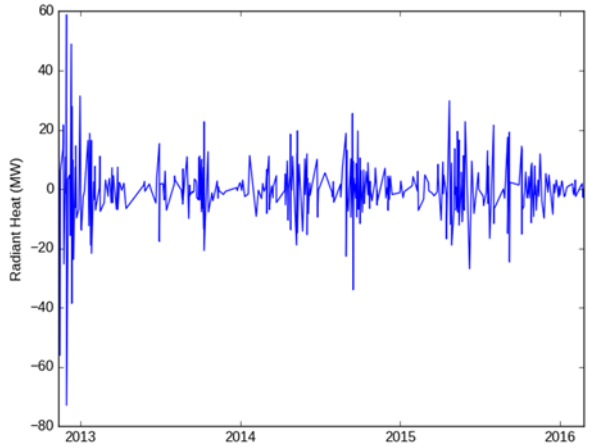


Figure 2 Chirpoi Snow Radiant Heat increment (MW) December 2012 - February 2016

Conclusion

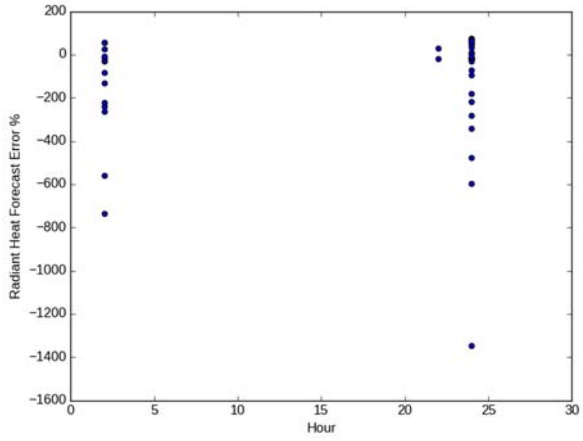


Figure 3 Relative error of ARIMA forecast

Unfortunately due to persistently poor weather conditions, there were too many gaps in data with a temporal distance between two readings coming up to several weeks. Thus, ARIMA did not manage to do reasonably accurate forecasts. As it could be seen on Fugure 3 relative error comes up to hundreds of percent. The situation could've been improved via introduction of new data sources, however many volcanoes are not easily accessible leaving little to no choice but to use satellite data only. For instance, Chirpoi volcano reviewed in this article is located at an uninhabited island. Moreover even if there does exist an instrumental network for a specific volcano it still requires a lot of work, both organizational and technical, to integrate its data, whereas satellite imagery serves as a universal data source. Hence the most reasonable approach seems to be to integrate data from different classes and generations of satellites to improve its temporal resolution and shorten the gaps between readings.

Future work

Satellite infrared imagery proves itself to be an interesting source of data for volcano research. Current plans involve, although not limited to:

1. data delivery latency improvements – via software installation at receiving station around the world (currently installed on Kamchatka);
2. data volume improvements – via incorporation of older satellites data and SAR satellites data;
3. better classification – via various clustering algorithms application and evaluation, preliminary results of a graph based hierarchical clustering have been very promising;
4. better insight – via neural networks based algorithms application, such an approach seem to be more stable in the presence of large data gaps.

This would not only allow better analysis of activity patterns in global volcanism, but will improve quality of neighboring data products such as the forest fire monitoring product.

References

[1] Christopher D. Elvidge et. al., VIIRS Nightfire: Satellite Pyrometry at Night, ISSN 2072-4292, http://www.mdpi.com/2072-4292/5/9/4423/pdf

[2] Andrew Harris, Thermal Remote Sensing of Active Volcanoes A USER'S MANUAL, ISBN 978-0-521-85945-5

[3] MIROVA Near Real Time Volcanic HotSpot Detection System http://www.mirovaweb.it

[4] MODVOLC Near-real-time thermal monitoring of global hot-spots http://modis.higp.hawaii.edu/

[5] Skipper Seabold, Jonathan Taylor, StatsModels statistical package, Josef Perktold, http://statsmodels.sourceforge.net/devel/index.html

[6] Curtis Seaman, Beginner's Guide to VIIRS Imagery Data, http://rammb.cira.colostate.edu/projects/npp/Beginner_Guide_to_VIIRS_Imagery_Data.pdf

[7] Grigory Trifonov (1), Mikhail Zhizhin (2), and Dmitry Melnikov, Nightfire method to track volcanic eruptions from multispectral satellite images, http://meetingorganizer.copernicus.org/EGU2016/EGU2016-5409-1.pdf

[8] Robert Wright, MODVOLC: 14 years of autonomous observations of effusive volcanism from space, http://www.higp.hawaii.edu/~wright/geol_soc_426.pdf

Подходы к агрегации данных и извлечению факторов в задаче поиска мошенничества в банковских транзакциях

© О. И. Травкин

Московский государственный университет имени М. В. Ломоносова,
Москва
travkin.o.i@gmail.com

Аннотация

В данной статье проводится обзор подходов к агрегации данных и извлечению факторов для исследования потребительского поведения клиентов в задаче поиска мошенничества в банковских транзакциях. Для выявления мошеннических операций с банковскими картами очень важно анализировать историческое потребительское поведение клиентов. В данной статье рассмотрено два подхода. Первый подход – на основе трёх профилей клиентов: глобального, локального и частотно-временного. Глобальный профиль строится с помощью кластеризации клиентов исходя из характеристик их транзакций, что позволяет более точно работать с новыми или неактивными клиентами. Локальный профиль – исходя из исторического потребительского поведения каждого клиента. Частотно-временной профиль – с помощью анализа паттернов в операциях клиентов, построенных на основе частот совершения транзакций за определённый промежуток времени. Второй подход – RFM (Recency-Frequency-Monetary). Его суть заключается в расчёте периодичности, частоты и объёма проводимых клиентом операций за определённый промежуток времени. Кроме этого, предлагается модификация алгоритма DBSCAN для частичного обучения, которая может позволить значительно улучшить точность результатов поиска мошенничества на основе выявленных профилей и RFM характеристик.

Работа частично поддержана РФФИ (гранты 14-07-00548, 16-07-01028).

1 Введение

Число пользователей банковских карт в России растёт стремительно. К сожалению, ещё более стремительно растёт количество мошенничества с картами. По данным компании FICO за 2013 год Россия является самой быстрорастущей страной по объёму потерь от мошенничества с банковскими картами [11].

Нетрудно увидеть, что проблема выявления мошенничества в банковских транзакциях стоит достаточно остро. Эффективный инструмент решения проблемы мошенничества – использование алгоритмов машинного обучения. Но для этого, в первую очередь, необходимо определить факторы, позволяющие выявлять мошеннические операции. Особенностью рассматриваемой области является то, что использование сырых данных (поток транзакций) не даёт приемлемого результата [5]. Поэтому данная работа будет сосредоточена на том, чтобы сделать обзор существующих подходов к агрегации и извлечению факторов из транзакционных данных. При таком подходе учитывается важная для выявления мошенничества информация о прошлом поведении клиента и его потребительских привычках.

Существуют различные алгоритмы машинного обучения, а также различные подходы к обучению. Для эффективного и качественного решения задачи выявления мошенничества необходимо выбрать такой подход и алгоритм, который наилучшим образом впишется в рассматриваемую область. Есть основания полагать, что наилучшим вариантом будет подход с частичным обучением. Он учитывают кластерную структуру неразмеченных данных, одновременно учитывая размеченные обучающие примеры. В статье будет приведено обоснование того, почему частичное обучение подходит для решения задачи мошенничества с банковскими картами наилучшим образом. Кроме того, будет предложена модификация алгоритма DBSCAN для выявления мошенничества в банковских транзакциях на основе алгоритмов частичного обучения, тестирование которой будет проведено в рамках следующих работ.

Дальнейшее изложение будет организовано следующим образом: в секции 2 будет дано описание предметной области, в секции 3 – приведён перечень методов, используемых для выявления мошенничества с банковскими картами. В секции 4 будет дан обзор подходов к агрегации данных и извлечению факторов. Далее, в секции 5 будет

Труды XVIII Международной конференции DAMDID/RCDL'2016 «Аналитика и управление данными в областях с интенсивным использованием данных», Ершово, 11-14 октября 2016

рассмотрен алгоритм на основе частичного обучения. В секции 6 приведены результаты эксперимента по применению алгоритма к симуляционным данным.

2 Предметная область

Мошенничество с банковскими картами принято делить на две большие категории: заявочное и поведенческое мошенничество. Заявочное мошенничество может возникнуть при получении новой карты в компании эмитенте [4]. Для предотвращения мошенничества такого типа часто используют кредитный скоринг. Что касается поведенческого мошенничества, то его делят на 4 типа: кража почты (ситуация, когда мошенник получает доступ к конверту с картой до того, как она придёт к законному владельцу), потерянные и украденные карты, подделанные карты (создание физической копии карты) и «без предоставления карты». Мошенничество «без предоставления карты», в отличие от трёх предыдущих не требует наличия самой карты – оно совершается удалённо с помощью украденной информации о реквизитах кары [5]. Это очень удобный способ мошенничества, так как он почти полностью анонимный.

Финансовые институты борются с мошенничества на двух уровнях: противодействие мошенничеству и выявление мошенничества [4]. Противодействие мошенничеству включает в себя все действия и меры направленные на то, чтобы мошенничество никогда не случилось. Сюда можно отнести активацию карты перед первым использованием, одноразовые пароли, пин-код и так далее. Что касается выявления мошенничества, то сюда относятся системы и практики по скорейшему выявлению мошенничества, если оно уже произошло [4]. Чем скорее мошенничество будет выявлено, тем меньше будут потери от него, так как карта и все транзакции по ней будут заблокированы.

В данной статье рассматриваются подходы к выявлению поведенческого мошенничества. Прежде всего, необходимо дать определение мошенничеству в рамках выбранной категории: операцию с банковской картой клиента будем называть мошеннической, если она совершается без ведома клиента и против его интересов. Кроме того, необходимо учитывать, что мы можем выявить только те мошеннические операции, которые реализовались достаточное число раз, не похожи по некоторым своим характеристикам на предыдущие операции клиента [15], но похожи на предыдущие выявленные мошеннические транзакции [18].

3 Методы выявления мошенничества

Наибольший интерес представляют статистические методы выявления мошенничества. Они делятся на две большие группы: обучение с учителем и обучение без учителя. Обучение без учителя использует характеристики клиента или его транзакции, чтобы разделить их на небольшие кластеры, максимально непохожие друг на друга. Если новая транзакция не попадает в один из кластеров, считающийся нормальным, то срабатывает триггер для такой транзакции [4]. В тоже время большинство работ рассматривает обучение с учителем, которое использует прошлые мошеннические транзакции, чтобы сделать вывод о подозрительности текущих. Наиболее распространённым инструментом в данной предметной области для обучения с учителем являются искусственные нейронные сети [2,6,9, 12,15,21,23], так как обычно с их помощью достигается более высокое качество. Тем не менее, получаемые модели не интерпретируемы. В последнее время также часто используются ансамбли методов, например, метод случайного леса [3,8,26]. К оставшимся методам, используемым для выявления мошенничества, можно отнести рассуждения на основе прецедентов [25], Байесовские сети [15], деревья решений [20], логистическую регрессию [3,21], скрытые Марковские цепи [22], ассоциативные правила [19], метод опорных векторов [3] и генетические алгоритмы [10].

4 Стратегия агрегации данных

При разработке моделей выявления мошенничества с кредитными картами, обычно, изначально доступен только сырой набор данных, включающий в себя исключительно информацию по индивидуальным транзакциям. В Таблице 1 перечислены атрибуты, которые присутствуют в большинстве выборок данных о транзакциях.

Таблица 1 Типичный состав атрибутов

Имя атрибута	**Описание**
Transaction ID	Уникальный идентификатор транзакции
Time	Дата и время транзакции
Account number	Идентификатор клиента
Card number	Идентификатор карты
Transaction type	Internet, ATM, POS…
Amount	Сумма транзакции
Currency	Валюта транзакции
Merchant code	Код вида торговой точки (MCC Code)
Merchant group	Группа торговой точки
Country	Страна проведения транзакции

Чаще всего сырых данных недостаточно для построения качественной модели [5], так как при выявлении мошенничества необходимо учитывать поведенческие особенности каждого клиента. Для этого создаётся набор новых переменных, включающих в себя информацию о предыдущих транзакциях. Для создания таких переменных применяют так называемую стратегию агрегации [26]. Данный подход наиболее популярен на текущий момент. Смысл агрегации – собрать информацию о транзакциях клиента за последнее время: сумму и

количество транзакций в разрезе карты, страны, валюты, вида торговой точки и так далее. Один из подходов к учёту поведенческих особенностей клиента основан на профилировании, когда для каждого клиента выделяются три составляющие: глобальный профиль, локальный профиль и частотно-временной профиль [18]. Кроме того, существует RFM подход, который позволяет разносторонне исследовать исторический потребительский профиль клиента с помощью расчёта периодичности, частоты и объёма транзакций в различных разрезах [24].

4.3 Глобальный профиль

Глобальное профилирование необходимо для того, чтобы определить глобальные группы клиентов, основанные на характеристиках их транзакций. Это поможет охарактеризовать тех, о ком не достаточно данных в обучающей выборке [18]. Алгоритм профилирования следующий. В первую очередь, для каждого пользователя рассчитываются: общее количество транзакций, средняя сумма транзакций, среднее время между двумя последовательными транзакциями, число транзакций из других стран, типичное (модальное) время (часовой интервал) транзакции. Кластеризация производится с помощью итеративной версии DBSCAN с использованием расстояния Махаланобиса [16]. Итеративная версия используется для того, чтобы избежать некоторых недостатков алгоритма DBSCAN, возникающих при применении на несбалансированной выборке с несколькими большими и множеством маленьких кластеров. Количество итераций и начальные значения для алгоритма подбираются эмпирически [18].

Обновление кластеров происходит с некоторой периодичностью, например, раз в месяц. Более частое переобучение почти не имеет смысла в виду того, что будет относительное небольшое количество данных и характеристики клиентов меняются не так быстро.

После того, как выборка разбита на кластеры, нам необходимо уметь быстро определять принадлежность новой заявки к тому или иному кластеру. Для этих целей можно использовать дерево решений, которое будет обучено на полученных кластерах на основе тех же факторов.

4.2 Локальный профиль

Локальный профиль необходим для того, чтобы охарактеризовать индивидуальные поведенческие закономерности в транзакциях клиента. Процесс локального профилирования состоит в аппроксимации эмпирического распределения атрибутов транзакций гистограммой [18]. Выбран именно этот способ ввиду его простоты, наглядности и эффективности.

В процессе работы алгоритма для каждой новой транзакции используется метод HBOS [13]. С его помощью рассчитывается вероятность операции в контексте предыдущих. Для этого строятся одномерные гистограммы по значениям каждого из атрибутов. Гистограммы нормируются так, чтобы максимальная высота столбца была равна единице. Чтобы оценить аномальность транзакции рассчитывается следующая величина для каждой транзакции в контексте оцененного распределения:

$$log \frac{1}{hist_i(t_i)}$$

где t_i – это i-ый атрибут транзакции t, $hist_i(t_i)$ – высота столбца гистограммы, в который попало новое значение. Если пришло значение фактора, которое не встречалось ранее для этого клиента, то для него в гистограмме используется частота, рассчитанная по его кластеру из глобального профиля. Если это не возможно, то используется частота, оценённая на всей выборке. Таким же образом действуем и с клиентами, данных по которым очень мало.

Выбор временного периода для расчёта эмпирического распределения очень важен [1]. Мы остановимся на 2-х временных интервалах: 7 дней – для учёта наиболее актуальных потребительских трендов и 4 недели – для выявления более долгосрочных потребительских особенностей.

Обновление частот гистограмм должно осуществляться с периодичностью, не большей чем минимальное временное окно. В нашем случае – это 7 дней. Идеальный вариант - ежедневно. Обновление происходит с использованием экспоненциального сглаживания. Это позволяет не делать расчёты со старыми данными, снижая нагрузку на вычислительную систему, и, в тоже время, позволяет учесть прошлую информацию о сделках в распределении факторов:

$$hist_i(t_i) = a * hist_i(t_i)_t + (1 - a) * hist_i(t_i)_{t-1},$$

где параметр a выбирается эмпирически.

Предлагается строить гистограммы для следующих атрибутов: номинальные – Merchant code, Merchant group, Country, Currency, Transaction type; интервальные – Amount, Time. Для номинальных переменных мы просто подсчитываем частоту появления каждой группы. Для интервальных – проводится процедура разбиения на интервалы. Для Time – оно переводится в часы, округляясь в большую сторону если минут больше чем 30 и в меньшую в противоположном случае, например: 21.03.2011 21:31 станет 22. Далее считается частота появления каждого часа. Для Amount – разбивается на 10 равномерных интервалов между максимальным и минимальным значением для клиента за период и подсчитывается число попаданий в каждый из интервалов [13].

4.3 Частотно-временной профиль

С помощью данного профиля выявляются мошеннические транзакции, которые маскируются под не мошеннические. Примером могут служить

небольшие по сумме, но очень частые транзакции, которые не будут пойманы с помощью локального или глобального профилей.

Таким образом, для каждого клиента рассчитывается абсолютная сумма транзакций в этот день, количество транзакций в день и максимальное количество транзакций в день за последнее время. Для каждого из этих факторов рассчитывается среднее значение и стандартное отклонение за выбранный период. Аномальной будет считаться та транзакция, которая выходит за интервал: среднее значение +/- стандартное отклонение. Для редко расплачивающихся картами клиентов данный профиль не создаётся. Обновление среднего значения также происходит с помощью экспоненциального сглаживания.

4.4 RFM

RFM подход основывается на расчёте периодичности, частоты и объёма транзакций клиента в различных разрезах и за различные периоды времени [24]. В первую очередь определим временные периоды. Возьмём такие же, как для построения локального профиля. Теперь определимся с разрезами. Наиболее удобным форматом для представления изучаемых разрезов и полученных факторов будет таблица, которая изображена далее.

Таким образом, мы видим, что RFM подход является аналогом частотно-временного профиля, но более детализированным. Кроме того, в рамках данного подхода выделяются факторы, характеризующие первичность транзакции в рассматриваемом временном интервале. Таким образом, оба подхода могут быть гармонично объединены в один, что должно повысить ранжирующие способности моделей, построенных на извлечённых факторах.

5 Частичное обучение

Одной из особенностей мошенничества с банковскими картами является то, что возможности банка по разметке обучающих данных сильно ограничены. Лишь небольшая доля сделок попадает в расследования специалистов противодействия мошенничеству, а клиенты не всегда сообщают о фактах мошенничества с их картами. Это приводит к тому, что в данных очень мало размеченных транзакций, а те что размечены имеют тенденцию быть мошенническими. Всё это снижает эффективность методов обучения с учителем. Кроме того, мошенники часто меняют свои стратегии вывода средств, чтобы оставаться непойманными, что снижает эффективность методов обучения с учителем ещё сильнее, так как алгоритмы этого типа наиболее чувствительны к тем мошенническим схемам, которые встречаются в выборке для обучения наиболее часто. Одним из решений описанных сложностей могло бы стать использование алгоритмов машинного обучения без учителя, для выявления аномальных транзакций или обнаружения кластеров в пространстве рассматриваемых признаков, но это приводит к другим, возможно более серьёзным проблемам. Без использования размеченных обучающих примеров, полученные результаты могут быть непредсказуемыми и тяжело трактуемыми. Именно поэтому алгоритмы обучения без учителя не получили широкого распространения в решении проблемы выявления мошенничества в банковских транзакциях. Наилучшим компромиссом в рассматриваемой ситуации будут алгоритмы с частичным обучением, которые находятся посередине между двумя упомянутыми ранее типами.

Таблица 2 Разрезы для расчёта факторов в рамках RFM

Recency	Время, прошедшее с предыдущей транзакции
MC	для данного вида торговой точки
MC Category	для данной категории торговой точки
Global	для всех транзакций клиента
Country	для всех транзакций в той же стране
Currency	для всех транзакций в той же валюте
Transaction type	для всех транзакций того же типа
Frequency	Общее количество транзакций
MC	для данного вида торговой точки
MC Category	для данной категории торговой точки
Global	для всех транзакций клиента
Country	для всех транзакций в той же стране
Currency	для всех транзакций в той же валюте
Transaction type	для всех транзакций того же типа
Monetary Value	Средняя сумма транзакции
MC	для данного вида торговой точки
MC Category	для данной категории торговой точки
Global	для всех транзакций клиента
Country	для всех транзакций в той же стране
Currency	для всех транзакций в той же валюте
Transaction type	для всех транзакций того же типа
Event occurrence	Первая покупка?
MC	для данного вида торговой точки
MC Category	для данной категории торговой точки
Global	для всех транзакций клиента
Country	для всех транзакций в той же стране
Currency	для всех транзакций в той же валюте
Transaction type	для всех транзакций того же типа

5.1 Основные принципы

Задача частичного обучения ставится следующим образом. Есть множество объектов X и множество классов Y. $X^i = \{x_1, \dots, x_l\}\{y_1, \dots, y_l\}$ – размеченная

выборка, $X^k = \{x_{l+1}, \dots, x_{l+k}\}$ – неразмеченная выборка. Необходимо построить алгоритм классификации a: $X \to Y$.

Существует несколько подходов к решению данной задачи. Первый и наиболее простой подход – это эвристические методы, такие как self-training и co-learning. Такие подходы требуют многократного обучения, поэтому вычислительно неэффективны. Второй – модификации методов кластеризации. Он достаточно прост в реализации (необходимо внести лишь некоторые ограничения), но, как правило, трудоёмкий в вычислениях. Наконец, модификации методов классификации. Данный подход реализуется сложнее, но даёт более вычислительно-эффективные методы.

5.2 Применение в области выявления мошенничества с банковскими картами

Применение методов частичного обучения в задачах поиска мошеннических транзакций весьма перспективна ввиду причин, перечисленных выше. Но, к сожалению, существуют ограничения, которые должны быть наложены на алгоритмы и в рамках частичного обучения. В ситуации, когда мы имеем практически один размеченный класс, выбор алгоритмов существенно ограничен: модификации алгоритмов классификации будут заведомо хуже или вовсе не применимы, так как для них необходима выборка хотя бы с двумя размеченными классами.

Прежде чем окончательно определиться с используемым алгоритмом, необходимо сделать предположение о структуре данных. Вероятнее всего, что мошеннические операции представляют собой небольшие скопления в пространстве признаков. Это обосновано тем, что мошенники часто действуют очень схожим образом: отрабатывают определённую работающую схему до тех пор, пока её не закроют, либо мошеннические действия совершает вредоносная программа на электронном устройстве клиента, которая действует по чёткому алгоритму. В тоже время поведение мошенников изменчиво, так как они находятся в постоянном поиске новых лазеек в системах противодействия и выявления мошенничества.

Исходя из представленных выше предположений, наилучшим будет тот алгоритм, который умеет выделять небольшие плотные скопления и ему не нужно задавать их количество. Одним из возможных вариантов являются алгоритмы на основе плотности, например, DBSCAN. В данной работе будет рассмотрена модификация данного алгоритма под поставленную задачу частичного обучения с одним размеченным классом.

Стоит отметить, что на текущий момент существуют реализации алгоритма DBSCAN с частичным обучением, например [14], но они не пригодны для применения в рамках описанной ранее предметной области, где преобладают наблюдения с одним размеченным классом.

5.3 DBSCAN

DBSCAN (Density Based Spatial Clustering of Applications with Noise) [17] – это алгоритм кластеризации, основанный на плотности и работающий следующим образом: пусть имеются точки в некотором пространстве, алгоритм объединяет вместе точки, находящиеся близко друг к другу (которые имеют много близкорасположенных соседей), а те точки, что лежат в областях с низкой плотностью (ближайшие соседи расположены далеко) оставляет в качестве шума.

Перейдём к модификации данного алгоритма для возможности использования частичного обучения с учётом ограничений, накладываемых предметной областью. Основные функции:

1. **expandCluster** (*D, D_l, P, NeighborPts, C, eps, MinPts*) – функция расширения кластера за счёт соседей, обладающих необходимыми параметрами *eps* и *MinPts.*
2. **trueEps** (*D, D_l, P*) – функция рассчитывающая *eps* и *MinPts* для оптимальной кластеризации.

Здесь *D* – все не размеченные точки, D_l – все размеченные точки, *P* – точка вокруг которой ищются соседи, *NeighborPts* – соседи рассматриваемой точки с необходимыми параметрами *eps* и *MinPts*.

В данной реализации поиск кластеров производится вокруг уже известных (размеченных) мошеннических наблюдений, что позволяет выявить другие потенциально мошеннические транзакции, а одна точка может принадлежать сразу нескольким кластерам. Смысл этих изменений в том, чтобы позволить алгоритму разбивать всё пространство на области, каждая из которых характеризуется некоторой мерой, отражающей шанс того, что в ней содержатся мошеннические транзакции. Причём некоторые из областей, вероятно, будут содержать в себе другие полученные области. Такой подход позволяет более тонко управлять процессом выявления мошенничества в зависимости от соотношения цены ошибки первого и второго рода.

Алгоритм выполняет следующие действия: для каждой размеченной точки он находит другую ближайшую размеченную точку и строит гиперсферу, диаметром которой является прямая, соединяющая две эти точки (eps'). С помощью гиперсферы мы оцениваем плотность точек в пространстве между двумя выбранными (подразумевается, что они принадлежат одному кластеру) для того, чтобы подобрать параметры DBSCAN, максимально отражающие структуру данных кластера. Параметру алгоритма *eps* присваивается значение равное радиусу сферы, умноженное на долю объёма гиперсферы, не заполненную точками:

$$0.5 * eps' * \left(1 - \frac{(V\ of\ n\text{-}cube) * len\left(distinct\left(scale\left(trueNeighbors\right)\right)\right)}{(V\ of\ n\text{-}sphere)}\right)$$

Рассмотрим выражение:

$$\frac{(V\ of\ n-cube)*len\left(distinct\left(scale\left(trueNeighbors\right)\right)\right)}{(V\ of\ n-sphere)},$$

где *trueNeighbors* – это число точек внутри гиперсферы, $(V\ of\ n-cube)$ – объём гиперкуба, построенного вокруг точки, $len\left(distinct\left(scale\left(trueNeighbors\right)\right)\right)$ – число уникальных точек, приведённых к необходимой шкале, $(V\ of\ n-sphere)$ – объём гиперсферы. Фактически, это отношение выражает долю объёма гиперсферы, заполненную точками. Процесс шкалирования – это преобразование коррдинат точек к такому виду, чтобы можно было заполнить гиперсферу неперсекающимися гиперкубами определённого размера, построенными вокруг точек. Например, заполним гиперсферу гиперкубами со стороной $k = eps'/100$, тогда, если мы преобразуем каждую координату как $round\left(x_i/k, 0\right) * k$, то сможем заполнить всё пространство внутри гиперсферы не персекающимися гиперкубами со стороной k. Таким образом, каждую уникальную точку (некоторые из них могли сойтись в одну из-за округления) мы заменяем гиперкубом и складываем их площади, получая оценку площади, занятой точками. Поделив это на объём гиперсферы, который можно приближённо вычислить (для n>3) по формуле: $V_n(R) \sim \frac{1}{\sqrt{\pi * n}} * \left(\frac{2\pi e}{n}\right)^{0.5*n} * R^n$, и вычитая из единицы, получаем долю объёма сферы, не заполненную точками. Умножая это на eps' получаем величину eps, которая тем больше, чем меньше заполнена гиперсфера. Этот же коэффициент применяем для определения оптимального *MinPts*: чем меньше заполнена сфера, тем меньше надо соседних точек для включения в кластер. Далее алгоритм, используя полученные параметры *MinPts* и *eps*, находит соседей двух рассматриваемых точек, объединяет их и действует по схеме функции **expandCluster**.

Таким образом, мы получили алгоритм, принимающий на вход две выборки (размеченные и не размеченные) и подбирающий все необходимые параметры для выявления кластеров автоматически. Это позволяет исключить практически все общеизвестные недостатки оригинального алгоритма DBSCAN: нет проблемы с граничными точками, так как фактически есть только один тип для кластера; значения параметров *MinPts* и *eps* подбираются автоматически; нет проблемы с различной плотностью кластеров, так как параметры подбираются индивидуально для каждого потенциального кластера. К сожалению, все эти изменения делают алгоритм более трудоёмким с точки зрения вычислений. Для того, чтобы облегчить вычисления, можно ограничивать возможные значения *eps*, так как в расмотрении очень больших областей смысла нет. Кроме того, есть потенциал в распараллеливании данного алгоритма и применении технологий Apache Spark [20].

5.4 Отбор факторов

Одним из важнейших этапов в процессе кластеризации является отбор факторов. Включение незначимых или сильно коррелирующих факторов может привести к тому, что адекватные кластеры не будут найдены. Существует два очевидных подхода по отбору факторов для кластеризации: использование априорных соображений и анализ важности факторов на специально подготовленной размеченной выборке. В нашем случае априорные соображения учтены полностью ввиду специфики подготовки факторов (подробнее об этом в секции 4). Что касается анализа на размеченной выборке, то этот способ кажется очень удобным в рамках поставленной задачи. Таким образом для отбора факторов предлагается использовать деревья решений. Это связано с необходимостью учитывать возможную нелинейность факторов относительно мошенничества, а деревья решений являются наиболее простым и понятным подходом для решения таких задач. Один из возможных вариантов реализации подхода к отбору факторов на основе деревьев решений заключается в следующем: для каждого фактора строится своё дерево решений, которое предсказывает известные факты мошенничества против всего остального. Предварительно отбираем обучающую выборку следующим образом: берём все размеченные данные, а потом дополняем их неразмеченными так, чтобы соотношение мошеннических транзакций к остальным было 1:1. Дальнейший алгоритм построен следующим образом: фактор разбивается на 40 равномерных интервалов (для интервальных факторов), каждому интервалу присваивается номер, и на этих номерах строится дерево решений, которое в итоге должно выдать не более чем 7 итоговых интервалов. Для номинальных переменных в дереве используются их фактические значения. Разбиение на 40 интервалов необходимо для того, чтобы обезопаситься от переобучения и получения очень маленьких итоговых интервалов. Для номинальных (категориальных) переменных также используется дерево решений, но без предварительного разбиения на интервалы. После этого рассчитывается коэффициент Gini для каждой переменной, отражающий способность фактора к ранжированию. Порог отсечения по Gini для факторов подбирается эмпирически. Полученный список значимых факторов проверяем на корреляцию и исключаем один из тех, по кому корреляция составила более 70%. Предпочтение отдаётся фактору, имеющему больший Gini. Таким образом, получаем финальный список факторов для кластеризации.

5.5 Обработка результатов

В результате применения описанного алгоритма, на выходе получаем набор кластеров, каждый из

которых характеризуется некоторой мерой, отражающей шанс того, в нём содержатся мошеннические транзакции, например, плотность кластера: отношение количества элементов в кластере к оценке объёма кластера. Кроме того, можно учитывать расстояние от точки до центра кластера, корректируя вероятность быть мошеннической для конкретной точки внутри кластера.

Теперь мы можем каждой новой транзакции поставить в соответствие выбранную меру, чтобы определить её шансы быть мошеннической (предварительно отнеся её к одному из кластеров). Далее, в зависимости от политики банка, применяются различные меры по борьбе с мошенничеством.

6 Результаты эксперимента

Результаты работы алгоритма были оценены на 10 симуляционных двумерных выборках. Двумерная выборка была выбрана в виду наибольшей наглядности результатов. Выборки были сформированы следующим образом. Вокруг трёх последовательно расположенных опорных точек были сформированы кластеры различной плотности. Кластеры формировались таким образом, чтобы точки внутри каждого кластера были распределены нормально по обеим координатам со средним значением равным соответствующей координате опорной точки. По такому же принципу в каждый из кластеров в небольшом количестве, пропорциональном его размеру, были добавлены размеченные (мошеннические) точки. После этого, равномерно по всему рассматриваемому пространству были добавлены точки шума, а также размеченные точки, не попадающие в кластеры. Пример на рисунке ниже. Укрупнённые точки – это размеченные (мошеннические) точки.

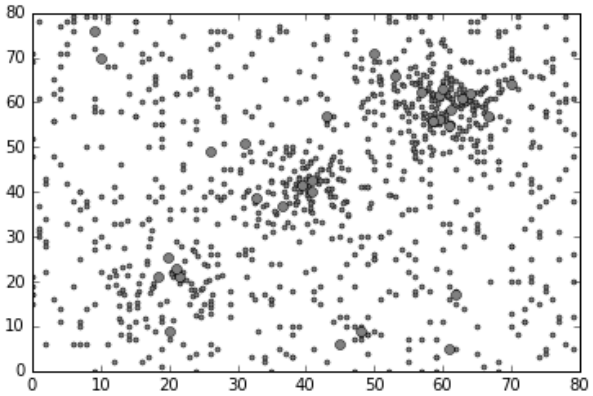


Рисунок 1 Пример симуляционной выборки

В рамках рассматриваемой задачи данный алгоритм будет применяться для классификации транзакций, поэтому в качестве метрики качества было решено использовать индекс Джини. Чтобы рассчитать данный индекс, предполагалось, что точки, находящиеся внутри кластера также являются мошенническими (данное предположение не использовалось для бучения, а только для расчёта качества). В результате оценивалось то, насколько точно и полно выделенные кластеры соответствуют реальным мошенническим кластерам.

Для того чтобы оценить качество разработанного алгоритма, результаты его работы были сравнены с результатами алгоритма HDBSCAN [7]. Его основные преимущества в том, что он применяет алгоритм DBSCAN к данным, используя различные значения параметра eps, подбирая лучшую кластеризацию, на основе стабильности данного параметра. Для работы алгоритм требует только один параметр – минимальный размер кластера. После применения алгоритма HDBSCAN так же оценивалось то, насколько точно и полно выделенные кластеры соответствуют реальным кластерам. Единственным исключением было то, что кластеры, которые не содержали изначально размеченные данные, исключались из анализа для повышения точности.

В результате была получена оценка среднего индекса Джини для алгоритма с частичным обучением: 85%, и для алгоритма HDBSCAN: 71,7%. Таким образом, мы видим, что предложенный подход оказывается в среднем лучше на 13 процентных пунктов.

Заключение

В данной статье произведён обзор подходов к агрегации данных и извлечению факторов в задаче поиска мошенничества с банковскими картами. Кроме того, обсуждалась предобработка данных с использованием деревьев решений и отбор факторов для оптимальной кластеризации, а также представлен новый алгоритм, являющийся модификацией DBSCAN для применения в задаче частичного обучения. Как показали эксперименты на симуляционных данных, данный алгоритм получает устойчивые положительные результаты на различных выборках без подбора параметров, необходимых классическому DBSCAN, кроме того, алгоритм показал себя лучше, чем алгоритм HDBSCAN.

В следующих работах будет произведено тестирование всех описанных подходов на реальных данных банковских транзакций.

Литература

[1] Alejandro Correa Bahnsen, Djamila Aouada, Aleksandar Stojanovic, Björn Ottersten. Feature engineering strategies for credit card fraud detection. Expert Systems With Applications 51 pp. 134–142, 2016.

[2] Aleskerov, E., Freisleben, B., Rao, B. Cardwatch: A neural network based database mining system for credit card fraud detection. In: Computational Intelligence for Financial Engineering (CIFEr),

1997., Proceedings of the IEEE/IAFE 1997. IEEE, p. 220–226, 1997.

[3] Bhattacharyya, S., Jha, S., Tharakunnel, K., Westland, J. C. Data mining for credit card fraud: A comparative study. Decision Support Systems 50 (3), p. 602–613, 2011.

[4] Bolton, R. J., & Hand, D. J. Statistical fraud detection: A review. Statistical Science, 17(3), p. 235–249, 2002.

[5] Bolton, R. J., & Hand, D. J. Unsupervised profiling methods for fraud detection. In Conference on credit scoring and credit control, Edinburgh, 2001.

[6] Brause, R., Langsdorf, T., Hepp, M. Neural data mining for credit card fraud detection. In: Proceedings. 11th IEEE International Conference on Tools with Artificial Intelligence. IEEE, p. 103–106, 1999.

[7] Campello R., Moulavi D., and Sander J. Density-Based Clustering Based on Hierarchical Density Estimates. In Advances in Knowledge Discovery and Data Mining, Springer. P. 160-172, 2013.

[8] Dal Pozzolo, A., Caelen, O., Le Borgne, Y.-A., Waterschoot, S., Bontempi, G. Learned lessons in credit card fraud detection from a practitioner perspective. Expert Systems with Applications 41 (10), p. 4915–4928, 2014.

[9] Dorronsoro, J. R., Ginel, F., Sgnchez, C., Cruz, C. Neural fraud detection in credit card operations. Neural Networks, IEEE Transactions on 8 (4), p. 827–834, 1997.

[10] Duman, E., Elikucuk, I. Solving credit card fraud detection problem by the new metaheuristics migrating birds optimization. In: Rojas, I., Joya, G., Cabestany, J. (Eds.), Advances in Computational Intelligence. Vol. 7903 of Lecture Notes in Computer Science. Springer Berlin Heidelberg, p. 62–71, 2013..

[11] FICO. Evolution of card fraud in Europe. Russia http://www.fico.com/landing/fraudeurope2013/country.php?countrycode=RUS

[12] Ghosh, S., Reilly, D. L. Credit card fraud detection with a neural-network. In: Proceedings of the Twenty-Seventh International Conference on System Sciences. Vol. 3. IEEE, p. 621–630, 1994.

[13] Goldstein, M., Dengel, A. Histogram-Based Outlier Score (HBOS): A Fast Unsupervised Anomaly Detection Algorithm, 2012.

[14] Levi Lelis, Jörg Sander. Semi-Supervised Density-Based Clustering. ICDM '09 Proceedings of the 2009 Ninth IEEE International Conference on Data Mining. p. 842-847. 2009

[15] Maes, S., Tuyls, K., Vanschoenwinkel, B., Manderick, B. Credit card fraud detection using Bayesian and neural networks. In: Proceedings of the 1st international naiso congress on neuro fuzzy technologies, 2002.

[16] Mahalanobis, Prasanta Chandra. On the generalised distance in statistics. Proceedings of the National Institute of Sciences of India 2 (1). p. 49–55, 1936.

[17] Martin Ester, Hans-Peter Kriegel, Jiirg Sander, Xiaowei Xu. A density-based algorithm for discovering clusters in large spatial databases with noise. Proceedings of the Second International Conference on Knowledge Discovery and Data Mining (KDD-96). AAAI Press. p. 226–231, 1996.

[18] Michele Carminati, Roberto Caron, Federico Maggi, Ilenia Epifani, Stefano Zanero. BankSealer: An Online Banking Fraud Analysis and Decision Support System. ICT Systems Security and Privacy Protection, Volume 428 of the series IFIP Advances in Information and Communication Technology, p. 380-394, 2014.

[19] S´anchez, D., Vila, M., Cerda, L., Serrano, J.-M. Association rules applied to credit card fraud detection. Expert Systems with Applications 36 (2), p. 3630–3640, 2009.

[20] Sandy Ryza, Uri Laserson, Sean Owen, Josh Wills. Advanced Analytics with Spark, Patterns for Learning from Data at Scale. O'Reilly, Pages: 276, 2015.

[21] Shen, A., Tong, R., Deng, Y. Application of classification models on credit card fraud detection. In: Service Systems and Service Management, 2007 International Conference on. IEEE, p. 1–4, 2007.

[22] Srivastava, A., Kundu, A., Sural, S., Majumdar, A. K. Credit card fraud detection using hidden markov model. Dependable and Secure Computing, IEEE Transactions on 5 (1), p. 37–48, 2008.

[23] Syeda, M., Zhang, Y.-Q., Pan, Y. Parallel granular neural networks for fast credit card fraud detection. In: Proceedings of the 2002 IEEE International Conference on Fuzzy Systems. Vol. 1. IEEE, p. 572–577, 2002.

[24] Veronique Van Vlasselaer, Cristian Bravo, Olivier Caelen, Tina Eliassi-Rad, Leman Akoglu, Monique Snoeck, Bart Baesens, APATE: A Novel Approach for Automated Credit Card Transaction Fraud Detection using Network-Based Extensions, Decision Support Systems, 2015.

[25] Wheeler, R., Aitken, S. Multiple algorithms for fraud detection. Knowledge-Based Systems 13 (2), p. 93–99, 2000.

[26] Whitrow, C., Hand, D. J., Juszczak, P., Weston, D., Adams, N. M. Transaction aggregation as a strategy for credit card fraud detection. Data Mining and Knowledge Discovery 18 (1), p. 30–55, 2009.

Data aggregation and feature extraction strategies for credit card fraud detection

Oleg Travkin

This paper provides an overview of approaches to data aggregation and feature extraction strategies for credit

card fraud detection. In order to identify credit card fraud, it is very important to analyze historical spending behavior of customers. This paper discusses two approaches. The first approach is based on global, local and temporal customer profiles. Global profile is built via clustering customers based on characteristics of transactions, that allows analyze new or inactive customers more accurate. Local profile is based on historical consumer behavior of each client. Temporal profile is based on analyzing patterns in customer transactions that are based on its frequency for a certain period of time. The second approach is called RFM (Recency-Frequency-Monetary). Within this approach the recency, the frequency and monetary volume of customer transactions are calculated for a certain period of time. In addition, we proposed semi-supervised modification of DBSCAN algorithm, which may allow to significantly improve the accuracy of modeling of fraud based on identified profiles and RFM characteristics.

Исследование методов поиска гендерных различий функциональной коннективности фМРТ покоя у здоровых людей среднего возраста

© С.И. Прийменко

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

mior12@mail.ru

Аннотация

В настоящее время в нейрофизиологии накопились большие наборы данных, полученные с помощью фМРТ. Несмотря на это, технологии для масштабируемого анализа больших объемов данных редко используются в данной сфере. В этой работе описаны методы для поиска и проверки гипотез о гендерных различиях функциональной связности фМРТ у здоровых людей среднего возраста, находящихся в состоянии покоя. Данные методы отличаются от стандартных методов тем, что в них используются различные способы уменьшения размерности пространства данных с потерей наименьшего количества информации. В работе также представлены данные проекта Human Connectome Project (HCP) и анализ формата nifti, в котором находятся эти данные. Предоставлен полный поток работ, который описывает все действия для нахождения связности участков головного мозга.

Работа частично поддержана РФФИ (гранты 14-07-00548, 16-07-01028).

1 Введение

В современном мире в разных областях науки наблюдается экспоненциальный рост данных [14, 19]. Для анализа больших объемов данных разработано множество специализированных инструментов, которые в первую очередь ориентированы на структурированные данные, но все чаще адаптированы и для более общих форм данных.

В том числе это применимо и в нейрофизиологии, где в настоящее время активно развивается направление нейровизуализации, которая позволяет визуализировать структуру, функции и биохимические характеристики мозга. В частности исследуются подходы, позволяющие находить связи отделов головного мозга [2]. Одним из направлений поиска связей в мозге является исследования людей, находящихся в состоянии покоя. В статье [15] рассматривается поиск функциональной связности состояния покоя у 24 мужчин и 17 женщин, участвовавших в военных действиях и получивших черепно-мозговую травму. В работе [18], проводились исследования для поиска различий функциональной связности 13 людей, страдающих болезнью Альцгеймера и 13 здоровых людей.

Функциональная связность – это связь между областями мозга, которые разделены функциональными свойствами. Она рассматривает отклонения от статистической независимости между распределенными и возможно пространственно удаленными нейронными единицами [13].

В работах [15, 18] анализировались относительно небольшие выборки данных, что является препятствием к обобщению полученных знаний. Кроме того, для выполнения таких задач не требуются большие вычислительные ресурсы и платформы. В работе [3] приводится всего 3 работы фМРТ (функциональная магнитно-резонансная томография), связанные с платформами для анализа больших объемов данных. С увеличением числа накопленных данных неминуемо потребуется планирование архитектуры программного обеспечения и выбор алгоритмов эффективных по памяти и времени выполнения.

Данная работа направлена на рассмотрение методов для поиска функциональной связности людей, находящихся в состоянии покоя, проекта HCP. Из-за большого количества данных, стандартные способы нахождения функциональных связей людей в состоянии покоя вычислительно затруднительны и нуждаются в огромных объемах памяти. Поэтому, также основной целью данной работы является нахождения аппроксимирующих методов, а также полное описание потока работ.

В разделе 2 описываются данные и вводятся необходимые определения. Далее в 3 представлен обзор методов для поиска и проверки гипотез. В разделе 4 описывается формат данных проекта HCP. В 5 приводятся поток работ. И в разделе 6

Труды XVIII Международной конференции DAMDID/RCDL'2016 «Аналитика и управление данными в областях с интенсивным использованием данных», Ершово, 11-14 октября 2016

подводится итог существующих путей решения нахождения функциональный связей мозга, а также формулируются дальнейшие шаги развития данной задачи.

2 ФМРТ покоя проекта HCP

В 2010 году стартовал проект Human Connectome Project [14]. Его основной целью является составление карты связности человеческого мозга с максимально возможной точностью. Всего в проекте планируется исследовать более 1200 здоровых взрослых людей и разместить эти данные в свободном доступе. На данный момент были получены фМРТ снимки для 900 человек в возрасте 20-35 лет. Далее в работе будут рассмотрены только данные ФМРТ покоя (rfMRI) [8].

Данные rfMRI были получены в результате проведения четырех сессий общей длиною 60 минут. В каждой из сессий пациент находился в затемненной комнате с открытыми глазами в расслабленном состоянии. После этого, полученные данные были очищены от различных шумов (движение головы, дыхание, сердцебиение) и нормированы. В результате данные представляют из себя около 900 наборов данных, в которых содержатся информация о пространственных и временной координатах и значениях вокселей для головного мозга [20].

Изображения были получены со следующими параметрами: период цикла съемки – 720 мс, толщина слоя – 2 мм и общее количество слоев – 72 штуки.

ФМРТ [9, 16] позволяет представить мозг в виде нарезки набора вокселей, где каждому вокселю соответствует временной ряд. ФМРТ показывает информацию об изменениях кровотока, которые сопровождают нейронную активность с относительно высоким пространственным разрешением, поэтому хорошо подходит для поиска функциональной связности [6, 21]. Этот процесс основан на связи мозгового кровотока и активности и называется BOLD-сигнал [2].

Одной из частых задач является вычисление функциональной связности между всеми элементами системы, независимо от того, есть ли прямые структурные связи этих элементов. Статистические закономерности между нейронными элементами колеблются от десятков или сотен миллисекунд.

В последние годы проводится большое количество исследований функциональных связей путем измерения разности между значениями вокселей временных рядов в разные моменты времени в различных областях мозга. Целью этих исследований является получение новых выводов о функциональных связях конкретных областей мозга [10]. В статье [15] рассматривается гипотеза о различии в паттернах орбифронтальной функциональной связности у женщин и мужчин-ветеранов с черепно-мозговыми травмами. В результате проведенного эксперимента эта гипотеза была подтверждена, так как женщины показали большую связность между левыми и правыми частями мозжечка и правой верхней теменной областью, а также между правой частью мозжечка и правой затылочной серединной областью. У мужчин была повышенная связность между левой и правой лобной частью и височной областью и увеличенная связность между правой лобной частью и левой островковой частью. Однако, из-за маленького размера выборки данная гипотеза нуждается в дальнейшей проверке на больших объемах данных. Одной из задач статьи является описать процедуру проверки описанной гипотезы на данных проекта HCP (здоровые люди среднего возраста, около 900 человек).

3 Обзор методов для поиска гипотез функциональной связности

3.1 Методы построения гипотез

Данные каждого субьекта могут быть представлены в виде матрицы $Y_{(t \times v)}$, где каждая строка представляет набор вокселей головного мозга в конкретный момент времени, а каждый столбец – временной ряд для соответствующего вокселя.

Перед тем как применить пространственное разложение методом независимых компонент (ICA) [1], к данным применяется метод главных компонент (PCA) или сингулярное разложение, представляющие данные так:

$$Y_{(t \times v)} \approx U_{(t \times n)} \times S_{(n \times n)} \times V^{T}_{(n \times v)},$$

где n – количество компонент, причем обычно это количество меньше, чем t, U - это набор временных собственных векторов, V - набор пространственных собственных векторов и S - диагональная матрица собственных значений. Метод главных компонент применяется к матрице V, оценивая новый набор пространственных отображений, которые являются линейными комбинациями отображения в V и максимально независимы друг от друга.

В случае анализа нескольких субъектов происходит построчная конкатенация всех s наборов испытуемых и применяется метод главных компонент, а затем метод независимых компонент, как описывалось выше. Полученный результат после использования метода главных компонент будет таким же приближением, как и выше, но теперь набор пространственных собственных векторов будет иметь размерность $n \times s$.

Данные фМРТ состояния покоя проекта Human Connectome Project требуют анализа, который проблематично выполнять из-за большого количества сложно-структурированных данных. В статье [17] представлено сравнение двух подходов для применения метода главных компонент на уровне групп: MIGP — MELODIC's Incremental Group-PCA [5] и SMIG — Small-Memory Iterative Group-PCA [11].

MIGP является итерационным подходом, который обеспечивает близкое приближение к

подходу с построчной конкатенацией данных с последующим применением метода независимых компонент, но без больших требований к памяти. Высокая точность достигается за счет сокращения наборов данных отдельных испытуемых с малым количеством главных компонент. Поэтапный подход сохраняет внутреннее PCA пространство *m* взвешенных пространственных собственных векторов, где *m* обычно больше, чем количество временных точек в каждом отдельном наборе. Под взвешенностью понимается то, что каждый пространственный собственный вектор умножается на соответствующее собственное значение. Окончательный набор *m* компонент может быть уменьшен до требуемой размерности *n* отбрасыванием остальных компонент.

SMIG – метод, который поворачивает матрицу данных для каждого субъекта с помощью вращения, полученного из корреляции исходной матрицы данных и матрицы данных усредненной группы. Для всех субъектов, повернутые матрицы могут быть усреднены, и PCA применяется без необходимости конкатенации данных, но необходимо совершить два прохода по всем исходным данным.

Оба метода дают близкое приближение к методу главных компонент, примененному к полной матрице, полученной в результате конкатенации по строке всех отдельных наборов данных. После применения этих алгоритмов к данным применяется метод независимых компонент и seed-based.

Seed-based [12] анализ функциональной связности между вокселями x_1 и x_2 можно определить как

$$C_{SB}(x_1, x_2) = \frac{\sum_{t=1}^{T} S(x_1,t) \times S(x_2,t)}{\sqrt{\sum_{t=1}^{T} S^2(x_1,t)} \sqrt{\sum_{t=1}^{T} S^2(x_2,t)}},$$

где $S(x, t)$ – аппроксимирует BOLD сигнал для вокселя x в момент времени t и T – количество временных точек в эксперименте. Знаменатель – коэффициент нормализации, который можно в данном случаи игнорировать.

BOLD сигнал можно разложить на компоненты (функциональные сети), каждый из которых содержит пространственный параметр и временную метку

$$S(x,t) = \sum_{k=1}^{K} M_k(x) A_k(t),$$

где K – число независимых компонент, M_k – пространственное отображение компоненты k-ой компоненты, A_k - временное отображение k-ой компоненты. В итоге, можно получить уравнение:

$$C_{SB}(x_1, x_2) = \frac{\sum_k M_k(x_1) M_k(x_2) \sum_{t=1}^{T} A_k^2(t)}{\sqrt{\sum_{t=1}^{T} S^2(x_1,t)} \sqrt{\sum_{t=1}^{T} S^2(x_2,t)}} + \frac{\sum_{k=l} \sum_{l=k} M_k(x_1) M_k(x_2) \sum_{t=1}^{T} A_k(t) A_l(t)}{\sqrt{\sum_{t=1}^{T} S^2(x_1,t)} \sqrt{\sum_{t=1}^{T} S^2(x_2,t)}},$$

где первое слагаемое является суммой в сети в пределах связности, а второе слагаемое является суммой между связностями.

$$C_{SB}(x_1, x_2) = Total\ WNC + Total\ BNC.$$

Для решения задачи поиска связности между различными удаленными компонентами возможно не учитывать значение Total WNC, тем самым сократить вычисление значения корреляции, не потеряв основную информацию о связях.

3.2 Статистические методы проверки гипотез

Для данной задачи нулевая гипотеза задается следующим способом: $H_0: \rho = 0$ коэффициент корреляции равен нулю для всех областей мозга. После выполнения seed-based корреляции, для статистической проверки гипотезы следует выполнить преобразование Фишера (так называемое Фишер z-преобразование), применяемый к коэффициенту корреляции.

Преобразование Фишера определяется как

$$z := \frac{1}{2} ln\left(\frac{1+r}{1-r}\right) = arctanh(r)$$

После использования преобразования Фишера на выходе получается величина, плотность распределения которой приближенно является гауссовой. В случае, когда значения входных данных близки к среднему, масштабирующий множитель близок к единице (участок графика для |X|<0.5). С другой стороны, для нормализованных значений на границах интервала масштабирующий множитель больше и выходные значения увеличиваются больше (участок для 0.5<|x|<1).

После преобразования Фишера, для каждого z проводится однохвостовый t-test и по нему принимаются или отвергаются гипотезы. Если значение p-value < 0.05, то нулевая гипотеза отвергается.

4 Данные

В качестве данных используются данные фМРТ состояния покоя (rfMRI) проекта Human Connectome Project [14].

Данные представлены в формате NIFTI [4]. В нем хранятся временные ряды вокселей в относительных координатах. Для получения мировых координат используется заголовок, в котором лежат необходимые данные для преобразования. В заголовке nifti-формата первые три измерения зарезервированы для определения трех пространственных координат, в то время как четвертое используется для определения моментов времени. Остальные измерения с 5 по 7 предназначены для других целей. Для совместимости, формат заголовка файла имеет размер 348 байт.

Поле dim_info хранит в одном байте направление кодирования частоты, направления кодирования фазы и направление разреза на слои.

Поле dim содержит размер матрицы изображения. Первый элемент содержит количество измерений (1-7). Если его значение не лежит в этом диапазоне, предполагается, что данные имеют обратный порядок байт. Второй, третий и четвертый элемент

отвечают за обозначения пространства (x, y, z), а пятый элемент отвечает за пространство времени t. Остальные размерности могут использоваться как угодно. Каждое i-ое поле (1-7) содержит положительное целое число, обозначающие длину i-го измерения.

Поле intent_code является целым числом, которое описывает данные. Некоторые требуют дополнительных параметров, которые содержатся в полях intent_p*, которые могут быть применены к изображению какое-то время или к 5-у измерению, если там хранится значение вокселя. В этом поле могут быть как статистические, так и нестатистические данные.

Поля полей slice_code, slice_start, slice_end и slice_duration полезны для хранения информации о тайминге fMRI и должны использоваться с полем dim_info, который содержит поле slice_dim. Если, и только если, поле slice_dim отлично от нуля, то slice_code имеет несколько возможных положений слоев.

Поле slice_start и slice_end сообщают, какой слой является первым, а какой последним, которые получаются после снимков МРТ. Все слои, присутствующие в изображении и не входящие в этот диапозон, рассматриваются как подбитые слои (обычно заполненные нулями). Slice_duration указывает количество времени, необходимое для обработки одного слоя.

Размер каждого вокселя хранится в pixdim[8] и каждый его элемент соответствует полю dim. Однако первый элемент имеет особое значение, которое равно либо 1 или -1. Информация про единицы измерения хранится в поле xyzt_units, который может быть закодирован для различных измерений.

Основное преимущество формата Nifti состоит в том, что он может хранить информацию об ориентации пространства. В стандартном файле предполагается, что каждая воксельная координата соответствует центру каждого вокселя. Мировые координаты системы являются следующими: x – ось, направленная поперек человеческой головы, y – ось, вдоль головы и z – вверх головы. Создатели данного формата разработали три метода для трансформации из воксельных координат в мировые. Первый метод поддерживает совместимость формата с форматом Analyze. Другие два метода используются для разных систем координат. В полях qform_code и sform_code задается нужный метод.

4.1 Метод 1

Мировые координаты определяются путем масштабирования по воксельной координате:

$$\begin{bmatrix} x \\ y \\ z \end{bmatrix} = \begin{bmatrix} i \\ j \\ k \end{bmatrix} \odot \begin{bmatrix} pixdim[1] \\ pixdim[2] \\ pixdim[3] \end{bmatrix},$$

где ⊙ - произведение Адамара.

4.2 Метод 2

Второй метод используется для указания координат сканера. Он также может быть использован для выравнивания изображения на предыдущей сессии одного и того же субъекта. Для компактности и простоты информация хранится в виде кватернионов (a, b, c, d), которые хранятся в полях quatern_b, quatern_c, quatern_d. Первый кватернион выражается через остальные по формуле: $a = \sqrt{1 - b^2 - c^2 - d^2}$. С помощью них строится матрица поворота:

$$R = \begin{bmatrix} a^2+b^2-c^2-d^2 & 2(bc-ad) & 2(bd+ac) \\ 2(bc+ad) & a^2+c^2-b^2-d^2 & 2(cd-ab) \\ 2(bd-ac) & 2(cd+ab) & a^2+d^2-b^2-c^2 \end{bmatrix}.$$

Эта матрица поворота вместе с размерами вокселя и сдвигом определяют мировые координаты:

$$\begin{bmatrix} x \\ y \\ z \end{bmatrix} = \boldsymbol{R} \begin{bmatrix} i \\ j \\ q * k \end{bmatrix} \odot \begin{bmatrix} pixdim[1] \\ pixdim[2] \\ pixdim[3] \end{bmatrix} + \begin{bmatrix} qoffset\_x \\ qoffset\_y \\ qoffset\_z \end{bmatrix},$$

где ⊙ – снова, произведение Адамара, а q = pixdim[0], равный либо 1, либо -1.

6.3 Метод 3

Использует аффинную матрицу, сохраненную в srow_*[4], которая отображает воксельный координаты в мировые:

$$\begin{bmatrix} x \\ y \\ z \\ 1 \end{bmatrix} = \begin{bmatrix} srow\_x[0] & srow\_x[1] & srow\_x[2] & srow\_x[3] \\ srow\_y[0] & srow\_y[1] & srow\_y[2] & srow\_y[3] \\ Grow\_z[0] & srow\_z[1] & srow\_z[2] & srow\_z[3] \end{bmatrix} \cdot \begin{bmatrix} i \\ j \\ k \\ 1 \end{bmatrix}.$$

Отличие от второго метода, который отображает воксельные координаты в мировые координаты сканера или выравнивает изображения одного и того же субъекта, 3 метод используется для преобразования в какое-то стандартное мировое пространство, например Talairach или MNI. В этом случае начало системы координат (0, 0, 0) находится на передней спайки мозга.

5 Поток работ

Для решения задачи поиска функциональной связности был специфицирован поток работ (см. Рис. 1). Далее описана каждая задача потока работ. Важным уточнением является, что данные предзагружены в HDFS [7].

1. loadData. Данные каждого субъекта загружаются в отдельный набор данных, где строки представляют значение вокселей в конкретный момент времени, а столбец значения соотвествующих вокселей во временном ряде.
2. splitGender. После загрузки данных, с помощью csv файла, который отдельно доступен в HCP, можно получить дополнительные факторы про каждого субъекта, такие как пол, возраст, доход, образование, IQ, вредные привычки и т.д. Для разделения объектов по половому признаку используется параметр возраст.

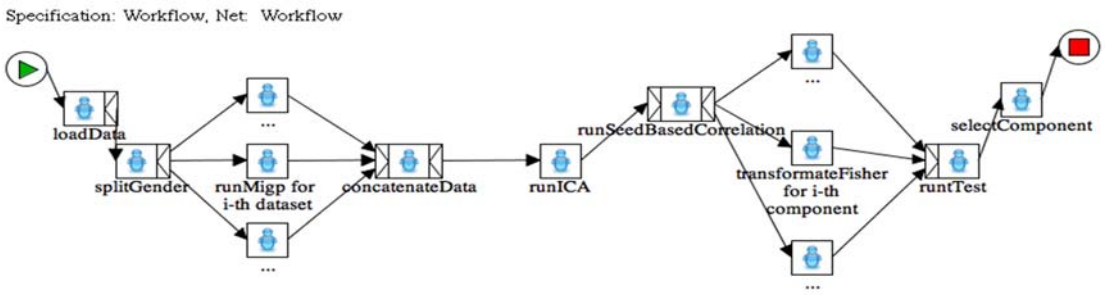


Рисунок 1 Описание потока работ

3. runMigp. Далее, отдельно для набора данных мужчин и женщин запускается алгоритм MIGP [5]. Этот алгоритм основан на итерационном применении PCA. Алгоритм выполняется параллельно в разных потоках, что позволяет эффективнее производить вычисления.
4. concatenateData. После завершения алгоритма MIGP, на выходе получается единственный набор данных, состоящих из *n* собственных векторов, имеющих наибольшие собственные значения.
5. runICA. Метод независимых компонент принимает набор собственных векторов, полученных на предыдущем этапе и возвращает вектора, которые максимально статистически независимы друг от друга.
6. runSeedBasedCorrelation. ICA алгоритм апроксимирует BOLD-сигнал, в связи с этим его можно разложить на пространственные и временную компоненты и рассматривать корреляцию только между внешними компонентами связей мозга.
7. transformateFisher. Для применения статистического теста, к полученным значениям корреляции применяется преобразование Фишера, который отображает интервал $(-1, 1)$ в интервал $(-\infty, +\infty)$.
8. runtTest. Статистический тест Стьюдента принимает z-value, полученные после преобразования Фишера и возвращает значение p-value.
9. selectComponent. После получения значения p-value, отбираются только те компоненты, для которых p-value < 0.05, тем самым отвергая нулевую гипотезу о том, что у данных областей мозга нет функциональный связей.

Заключение

В данной работе описаны методы для построения гипотез гендерных различий функциональной связности данных фМРТ состояния покоя проекта HCP. Для вычисления функциональной связности вначале используется метод корреляции seed-based. В силу большого количества данных проекта HCP перед корреляцией необходимо уменьшить размерность пространства данных. В статье рассмотрены методы MIGP и SMIG, которые уменьшают размерность данных таким образом, чтобы было потеряно наименьшее количество информации. Для проверки гипотез приводится обзор классических статистических подходов. В статье также описаны основные поля nifti-формата и работа с ними, составлен и представлен поток работ с кратким описанием каждого шага.

В дальнейшем планируется разработать архитектуру программного обеспечения, провести эксперименты на данных проекта HCP для проверки гипотезы, представленной в статье [15], и определить гендерные различия функциональной связности здоровых людей между разными регионами мозга.

Благодарность

Автор статьи выражает благодарность Н.В. Пономаревой и Д.Ю. Ковалеву за предоставленную идею.

Литература

[1] Christian F. Beckmann, Marilena DeLuca, Joseph T. Devlin and Stephen M. Smith. Investigations into Resting-state Connectivity using Independent Component Analysis. Philosophical transactions of the Royal Society of London, 360(1457), May 2005, p. 1001-1013.

[2] B.B. Biswal, J. Van Kylen, J.S. Hyde. Simultaneous assessment of flow and BOLD signals in resting-state functional connectivity maps. NMR in Biomedicine, 1997, 10, p.165-170.

[3] Roland N. Boubela, Klaudius Kalcher, Wolfgang Huf, Christian Našel and Ewald Moser. Big Data Approaches for the Analysis of Large-Scale fMRI Data Using Apache Spark and GPU Processing: A Demonstration on Resting-State fMRI Data from the Human Connectome Project. Frontiers in Neuroscience, 9, 2015. http://journal.frontiersin.org/article/10.3389/fnins.2015.00492/full

[4] Hester Breman. Diagrams of the NIfTI-1.1 file structure. http://nifti.nimh.nih.gov/nifti-1/documentation/nifti1diagrams_v2.pdf.

[5] V. Calhoun, T. Adali, G. Pearlson and J. Pekar. A method for making group inferences from functional MRI data using independent component analysis. Hum. Brain Mapp.;14(3), 2001, p. 140–151.

[6] K. J. Friston. Functional and Effective connectivity: a review. Brain Connectivity, 1(1), June 2011.p. 13-36.

[7] HDFS Architecture Guide. https://hadoop.apache.org/docs/r1.2.1/hdfs_design.html.

[8] M.P. van den Heuvel, H.E. Hulshoff Pol. Exploring the brain network: A review on resting-state fMRI functional connectivity. European Neuropsychopharmacology, August 2010, 20(8), p. 519-534.

[9] S. A. Huettel, A. W. Song, G. McCarthy. Functional Magnetic Resonance Imaging . Sunderland, MA: Sinauer Associates, Inc, 2- edition, 2009.

[10] Alexander G. Huth, Wendy A. de Heer, Thomas L. Griffiths, Frederic E. Theunissen and Jack L. Gallant. Natural speech reveals the semantic maps that tile human cerebral cortex. Nature International weekly journal of science, 532(7600), April 2016, p. 453-458.

[11] Hyvärinen A., Smith S. Computationally efficient group ICA for large groups. In Proceedings of Annual Meeting of the Organization for Human Brain Mapping. 2012.

[12] Suresh E. Joel, Brian S. Caffo, Peter C.M. van Zijl and James J. Pekar. On the relationship between seed-based and ICA-based measures of functional connectivity. Magnetic Resonance in Medicine, 66(3), September 2011, p. 644–657.

[13] A. Martin Lindquist. The Statistical Analysis of fMRI Data. Statistical Science, 23(4), May 2009, 439-464. http://projecteuclid.org/euclid.ss/1242049389

[14] D. S. Marcus, J. Harwel, T. Olsen, M. Hodge, M. F. Glasser, F. Prior and D. C. Van Essen. 2011. Informatics and data mining tools and strategies for the human connectome project. Front. Neuroinform.5. Available at: http://www.ncbi.nlm.nih.gov/pmc/articles/PMC3127103/ (accessed February 10, 2015).

[15] E. McGlade, J. Rogowska and D. Yurgelun-Todd D. Sex differences in orbitofrontal connectivity in male and female veterans with TBI. Brain Imaging and Behavior, 9(3), September 2015, p 535–549.

[16] A. Peter Rinck. Magnetic Resonance, a critical peer-reviewed introduction; functional MRI. In Proceedings of European Magnetic Resonance Forum. November 2014.

[17] S.M. Smith, A. Hyvärinen, G. Varoquaux, K.L. Miller,C.F. Beckmann. Group-PCA for very large fMRI datasets. Neuroimage, November 2014, 101, p. 738-749.

[18] Liang Wang, Yufeng Zang, Yong He, Meng Liang, Xinqing Zhang, Lixia Tian, Tao Wu, Tianzi Jiang and Kuncheng Li. Changes in hippocampal connectivity in the early stages of Alzheimer's disease: Evidence from resting state fMRI. Neuroimage,31(2), June 2006, p. 496-504.

[19] Large Synoptic Survey Telescope (LSST) project. Opening a window of discovery of the dynamic universe. http://www.lsst.org.

[20] WU-Minn HCP 900 Subjects Data Release: Reference Manual. December 2015. http://www.humanconnectome.org/documentation/S900/HCP_S900_Release_Reference_Manual.pdf.

[21] Ю.А. Селивёрстов, Е.В. Селивёрстова, Р.Н. Коновалов, М.В. Кротенкова, С.Н. Иллариошкин. Функциональная магнитно-резонансная томография покоя: возможности и будущее метода. Анналы клинической и экспериментальной неврологии, 7(4), 2013, стр. 39-44, Москва.

Research methods to search for the gender differences of functional connectivity of rest state fMRI in healthy middle-aged people

Sergey Priymenko

There are a lot of big data sets in neuroimaging that have been collected by fMRI nowadays. Despite of that, people barely use Large-Scale algorithms to analyze such data. This article introduces the number of methods to investigate a difference among functional connectivity of resting state fMRI by gender for healthy middle-aged people. These methods differ from conventional methods in that they use different approaches for reducing the dimensionality data space thus wasting the smallest amount of information. The article also presents the Human Connectome Project Project (HCP) and the format of the analysis nifti, which contains the data. It describes a workflow term that describes all the steps to find connectivity parts of the brain.

Использование СУБД Динамической Информационной Модели для анализа и обработки данных

© А. Н. Петров

Ярославский государственный университет им. Демидова,
г. Ярославль
axel_petroff@mail.ru

Аннотация

Рассматриваются возможности использования механизмов СУБД Динамической Информационной Модели (DIM) для анализа и обработки данных. Даются обобщенные сведения о способах организации и связывания данных DIM, а также средствах обработки данных: специализированных языках программирования PL\ODQL, DIM-FL, DIM-Script и механизме взаимодействий. Исследование расширяет и дополняет PL\ODQL механизмами работы с множествами и индексами. Приводится теоретическое обоснование алгоритмической полноты механизмов изменения свойств данных в СУБД DIM. В заключении рассматриваются разработанные программные компоненты инфраструктуры DIM, а также среда программирования приложений DIM – «DIM Developer» и варианты её использования для разработки сторонних приложений на основе DIM.

1 Динамическая Информационная Модель

Анализ данных является наиболее одной из важнейших задач информатики, имеющей глубокое прикладное значение в различных отраслях науки и производства. Поскольку большие объемы данных, доступные для многостороннего анализа консолидируются в базах данных, именно на их основе строятся системы анализа и системы поддержки принятия решений [1].

В настоящее время на рынке присутствует большое количество СУБД различных специализаций, которые можно разделить на несколько основных категорий:

- реляционные;
- объектно-ориентированные;
- объектно-реляционные;
- темпоральные СУБД;
- иерархические;
- сетевые;
- СУБД «ключ-значение»;
- «in-memory» СУБД.

Однако во многих случаях они являются OLTP-системами или же хранилищами данных и обладают существенными недостатками, не позволяющими эффективно их использовать в качестве средств анализа данных.

В стремлении преодолеть недостатки существующих систем была поставлена задача разработки и реализации нового объектного подхода к созданию СУБД, предполагающего более сложную систему связей между объектами и их типами, с возможностью изменения не только самих объектов, но их схемы типов. Разрабатываемая СУБД была названа *динамической информационной моделью* (*DIM*) [2].

В этом подходе было выделено 6 базовых отношений объектов: *наследования*, *включения*, *внутреннего наследования*, *внутреннего включения*, *истории* и взаимодействия. А также были разработаны объектный язык запросов ODQL и другие механизмы программирования для описания динамики данных. Такая развитая структура отношений между данными облегчает описание взаимосвязей между объектами, а также расширяет возможности поиска новых зависимостей между данными. Введение новой технологии СУБД потребовало обоснования статической полноты описания данных и полноты динамики данных этой модели. Последнее связано с точным описанием предметной области, а если быть более точным, с созданием ее математической модели, которая была названа *OD*-моделью.

Использование разработанного и реализованного языка запросов ODQL позволяет достаточно эффективно выполнять поиск данных, составляя запросы более короткие и понятные, чем на аналогичном языке для реляционных СУБД – SQL [3].

2 Программирование логики обработки данных в DIM

Кроме возможностей выполнения запросов поиска данных, СУБД *DIM* обладает широкими

Труды XVIII Международной конференции DAMDID/RCDL'2016 «Аналитика и управление данными в областях с интенсивным использованием данных», Ершово, 11-14 октября 2016

возможностями программирования логики приложений с использованием специализированных языков программирования: PL\ODQL [4] – языка общего назначения и DIM-FL [5] – языка математических формул для кодирования регламентированных расчетов. Использование нескольких языков программирования и унифицированного интерфейса вызова программных модулей позволяет более эффективно разделять труд программистов и инженеров, активнее использующих математические формулы при разработке прикладной бизнес-логики.

Более сложные механизмы динамического анализа и обработки данных реализованы с помощью механизма *взаимодействий*.

Взаимодействие – это инкапсуляция алгоритма обработки данных модели в *DIM*, его составляющие определяют алгоритм следующим образом:

Откуда – объект или группа объектов в момент t, у которых изменяются значения свойств;

Куда – объект или группа объектов, получающая в момент $t+1$ новые значения свойств;

Что – свойство или группа свойств-объектов, у которых изменяются значения этих свойств;

Как – процедура изменения свойств (тело взаимодействия).

2.1 Язык PL\ODQL и множества с индексами

В целях эффективной организации процедур для выполнения *взаимодействий* при разработке модели *DIM* были введены «*множества*» объектов, классов, свойств и их значений. Но они являются не вполне множествами, так как, с одной стороны, элементы этих «*множеств*» могут повторяться, как в мультимножествах, а с другой стороны, в некоторых случаях они обрабатываются как списки. При использовании сущностей в качестве элементов множеств, их объектный идентификатор будет выступать в роли главного индекса. Ранее эти особенности не были учтены, а потому не были организованы удобные операторы языка для работы с ними. В последних исследованиях были ликвидированы указанные пробелы, и упорядочено описание языка PL\ODQL в целях дальнейшей эффективной реализации взаимодействий.

Для описания типов различных множеств объектов введён PL\ODQL тип *множество*:

TYPESET ‘<’<тип множества>’>’:=ODQL запрос

Каждый элемент множества такого типа содержит свойства, описываемые фразой **SELECT** запроса (для одного или нескольких связанных классов).

Для обеспечения возможности работы с индексированными списками элементов множеств добавляется тип PL\ODQL *множества с индексами*:

[**CONST**] **SET**
{<тип множества>
| <сущностный тип>
| <простой тип>}
<имя множества>[:=<набор элементов>]
[**INDEX** <список индексов>],

где список индексов представляется в виде:

<список индексов>::=<индекс>[,<индекс>]…
<индекс>::=<имя индекса>
:<выражение от свойств типа индекса>,

где <имя индекса> подчинено тем же правилам именования, что и идентификаторы переменных PL\ODQL.

Данная конструкция позволяет присваивать переменной, имеющей тип **SET**, множество значений простого типа, или уже описанного типа множества, или множество значений сущностей системы DIM, которое будет получено в результате выполнения ODQL-запроса. В результирующем множестве могут присутствовать только значения указанного простого типа, или типа множества, или сущностного типа. При включении (добавлении) элемента во множество он получает по умолчанию индекс (главный), равный объектному идентификатору класса, объекта или свойства объекта в соответствии с сущностным типом множества, если источником элемента был запрос, или равный порядковому номеру включения элемента во множество, если источником элементов был список элементов. Непосредственно значение индекса элемента нельзя получить, так как он является указателем на элемент, и значение такого указателя зависит от реализации. Но в цикле **FOREACH,** где перебираются все элементы множества, можно запомнить это значение в индексной переменной, которая объявляется следующим образом:

VAR index <имя индексной переменной>
[:=<параметр цикла **FOREACH**>
| <имя индекса>]

Следовательно, каждый элемент множества при его создании получает индекс, с помощью которого к нему можно получить доступ следующим образом:

<имя множества>[<имя индексной переменной>]

Получив таким образом доступ к элементу множества, появляется также возможность удаления элемента присвоением ему значения **NULL** в рамках индекса:

<имя множества>[< имя индексной переменной >]
=**NULL**.

Введённый также в язык цикл **FOREACH** позволяет перебирать элементы множеств в отсортированном порядке, заданном индексом, что упрощает доступ к необходимым элементам.

Для манипуляции элементами множеств в язык PL\ODQL введены операции добавления – **ADD** и удаления – **EXCL** элементов, а для манипуляции самими множествами – бинарные операции +, *, −, которые возвращают, соответственно, значение множества объединения, пересечения и разности множеств операндов. Дополнительно вводятся операции сравнения множеств: ‘=’ – равенства и ‘<=’ – включения, а также операция принадлежности элемента множеству – **in**. При выполнении операций над множествами с сущностными элементами элементы определяются главным индексом, а при

операциях над множествами с элементами-значениями операции могут быть корректно выполнены над множествами с одинаковым типом элементов и только если для множеств задан индекс, позволяющий однозначно определить элемент множества.

Использование операций над большими множествами объектов хранимых в БД позволяет более эффективно и наглядно реализовывать задачи определения соответствий данных критериям, а упорядочивание и индексация данных – быстрее получать нужный результат при их анализе.

2.2 Проектирование и использование взаимодействий

Как уже было описано, *взаимодействия* инкапсулируют алгоритмы обработки данных в DIM, при этом управляющие структуры размещаются в *теле взаимодействия*.

Тело взаимодействия – основа алгоритма обработки данных, описывается на языке DIM-Script, и может включать в себя вызовы модулей *DIM*, написанных на специализированных языках PL\ODQL и DIM-FL – языке математических формул.

Использование единого интерфейса выполнения взаимодействий позволяет повсеместно применять их в *DIM* для выполнения различных действий. К примеру, создание пользователя реализовано как вызов взаимодействия **USER.CREATE**, при этом в качестве объекта *Что* выступает создаваемый пользователь, объекта *откуда* – инициирующий создание пользователь, а объекта *куда* – сама DIM, где и сохраняется информация о создаваемом пользователе.

Также следствием однообразия интерфейсов взаимодействий является упрощение механизма журналирования их вызовов, что позволяет упростить поиск «узких мест» в процессах обработки данных.

Разработанный механизм, однако, не позволял расширять правила обработки данных, не прибегая к написанию новых взаимодействий, и к тому же лишал пользователя возможности параллельной обработки данных. Чтобы избавится от этих недостатков, концепция *взаимодействий* была расширена до концепции системы графов взаимодействий, являющихся по сути «workflow»-процессами, наглядно описывающих реальные. Такой подход позволяет консолидировать процессы обработки данных в БД и избавиться от необходимости написания программного кода, реализующего процессы обработки данных, пользуясь лишь методами визуального проектирования, являющегося во многих случаях наиболее продуктивным.

Продолжая приведённый ранее пример с созданием пользователя, можно достроить цепочку взаимодействий, добавив после вызова собственно создания пользователя параллельное выполнение 2-х взаимодействий, а, следовательно, параллельное выполнение запросов, что позволяет ускорить общее время выполнения набора взаимодействий: активации пользователя **USER.ACTIVATE** и, к примеру, присвоение ролей созданному пользователю **USER.ADD_ROLES**.

3 Динамическая полнота DIM

Разработанный аппарат программирования бизнес-логики в виде взаимодействий может применяться для формулирования любых алгоритмов и переноса их из других систем. Однако, такое утверждение требует обоснования алгоритмической полноты механизмов изменения свойств данных в СУБД DIM, гарантирующей возможность переноса реализаций алгоритмов.

Ранее была сформулирована и доказана теорема о статической полноте СУБД *DIM* [6]:

Произвольная OD-модель для произвольного момента времени $t \in T$ *может быть статически описана с помощью некоторой схемы* S *классов DIM, находящейся в нормальной форме.*

Доказательство данной теоремы обосновывает возможность переноса данных из других систем и схем хранения данных в объекты *DIM*. Для доказательства статической полноты *DIM* в ранних работах было построено отображение G произвольной *OD*-модели в *схему классов DIM*.

В рамках исследования задача продолжить это отображение на алгоритмы множества F *OD*-модели во взаимодействия *DIM* таким образом, чтобы оно осталось согласованным с объектами обоих моделей, их свойствами и значениями свойств перед непосредственным выполнением каждого алгоритма $f \in F$ (соответствующего взаимодействия $G(f) \in B$) и сразу после его выполнения. Чтобы отображение $G(f)$ возможно меньше зависело от конструкций описаний f и $G(f)$, прибегнем к универсальному описанию алгоритма в виде *машины Тьюринга* (МТ) и для этого опишем *OD*-модель МТ (*OD*.МТ) и *DIM*-модель МТ (*DIM*.МТ).

По построенной МТ для *OD*-модели строится отображение $G(OD.MT) = DIM.MT$, которое позволяет построить МТ *DIM* путем преобразования функциональной таблицы *OD*.МТ в *DIM*.МТ, а также структуры входных и выходных состояний ленты обеих МТ.

С вызовом в некоторый момент $t \in T$ произвольного алгоритма $f \in F'$ произвольной *OD*-модели связаны 2 её статических описания при помощи схем классов *DIM*: S_0 в момент t непосредственно перед вызовом f и S_1 в момент $t+1$ завершения выполнения f. Будем говорить, что *некоторое взаимодействие DIM* $b \in B$

описывает динамику значений свойств объектов, вызываемую алгоритмом f, *если оно преобразует схему классов DIM* S_0 *в схему классов* S_1.

Теорема. *О полноте динамики значений свойств объектов*. [7]

Для произвольного алгоритма $f \in F'$ *произвольной OD-модели существует взаимодействие DIM* $b \in B$, *описывающее динамику значений свойств объектов, вызываемую алгоритмом* f *в произвольный момент* $t \in T$.

Доказательство: Пусть, произвольный алгоритм f изменения данных объектов *OD*-модели представлен в виде *машины Тьюринга OD.MT*. Для доказательства теоремы необходимо построить *взаимодействие DIM*, выполняющее те же модификации данных объектов DIM, что исходный алгоритм f над объектами *OD*-модели. Отделим функциональность МТ от исходных объектов и перенесём функции *OD.MT* в *DIM.MT*, это позволит выполнить операции *OD.MT* в рамках *DIM* и над объектами *DIM*.

1. построим модель машины Тьюринга в рамках *DIM*, выполняющую те же модификации данных объектов DIM, как и объектов *OD*-модели. Как было описано выше, *DIM.MT* может быть смоделирована в виде специальной процедуры, преобразующей набор символов, находящихся на ленте памяти. Исходя из того факта, что логика функционирования *OD.MT* определяется набором функций f_d, f_b и f_q, зависящих от текущего состояния МТ и текущего обозреваемого на ленте символа, становится возможным эмулировать выполнение этих функций в рамках *DIM.MT* внутри модуля DIM_MT в виде функций f_d^{DIM}, f_b^{DIM} и f_q^{DIM}.

2. в *DIM*, взаимодействия принимают в качестве параметров набор объектов, получаемых в результате определения параметра *Откуда*, в отличие от объекта ленты памяти, используемой *DIM.MT* в качестве параметра;

3. таким образом, необходимо построить преобразование между набором объектов *DIM* и объектом ленты памяти o_{Lenta}^{DIM} (включающем в себя массив объектов o_{Cell}^{DIM}, используемый как лента памяти, следовательно, $i_{\min}$ всегда равняется 0, а $i_{\max}$ равняется длине массива объектов o_{Cell}^{DIM}). Подобная трансформация может быть выполнена специальной кодирующей функцией, «записывающей» данные объектов на «ленту памяти» согласно формату заданному *OD.MT* (формат представления объектов на ленте памяти является предопределенным). После построения объекта ленты памяти, удовлетворяющего условиям $G^{-1}(o_{DIM}^{Lenta}) = o^{Lenta}$ и $G^{-1}(o_{i_{DIM}}^{Cell}) = o_i^{Cell}, \forall i \in [i_{\min}, i_{\max}]$, можно переходить к построению основного цикла процедуры, моделирующей выполнение МТ;

```
BEGIN
// основной цикл
WHILE (q <> q0)
BEGIN
d := dim_fd(get(tape, i), q); // сдвиг
b := dim_fb(get(tape, i), q); // записываемый символ
q := dim_fq(get(tape, i), q); // новое состояние
i := i + d;                   // текущее состояние
. . .
```

4. после останова МТ (после завершения итераций основного цикла) необходимо произвести обратную трансформацию: преобразовать объект ленты памяти *DIM.MT* обратно в объекты DIM.

Полученный в результате трансформации набор объектов является результатом работы *OD.MT* над объектами *DIM*. Однако, в объектах, полученных таким образом, не всегда корректно отражаются изменения, приобретённые ими в результате применения *OD.MT* к объектам *DIM*. К примеру, если был модифицирован атрибут объекта, принадлежащий родительскому объекту (унаследованный атрибут), попытка сохранить его в БД приведёт к изменениям других объектов, являющихся дочерними к объекту, от которого унаследован атрибут, что является некорректным поведением. Решение состоит в том, что в случае, когда не все дочерние объекты были изменены, родительский объект копируется, модифицируется и становится родительским для объектов, атрибуты которых были изменены в процессе выполнения МТ. В таком случае объекты, атрибуты которых не были явно изменены, унаследуют корректное значение атрибута.

Реализация такого решения – это применение процедуры *NORMALIZE*, использующей 2 множества объектов: *исходное* множество объектов и *результирующее*:

1. для каждого объекта o_s из множества исходных объектов найти соответствующий ему объект o_r из результирующего множества;

2. если исходный объект o_s был модифицирован, определить, принадлежат ли измененные атрибуты самому объекту, либо они унаследованы;

3. если были изменены унаследованные атрибуты, необходимо создать копию первоначального объекта в результирующем множестве, а для объекта o_r - изменить связь к

родительскому объекту на связь к неизмененной копии родительского объекта.

Таким образом, значения атрибутов, которые не были изменены в процессе работы МТ, не будут нарушены.

Аналогично, подобные операции необходимо провести с объектами включения и включающими объектами – в данных случаях, создаётся новый объект взамен изменяемого. Это реализовано в виде специальной процедуры *NORMALIZE_INCLUSION*, использующей 2 множества объектов: *исходное* и *результирующее*:

1. для каждого объекта o_s из множества исходных объектов, являющегося объектом включения или включающим объектом, найти соответствующий ему объект o_r из результирующего множества;
2. если исходный объект o_s был модифицирован, создать его копию в результирующем множестве $o_{r_{copy}}$;
3. связать отношением *истории* объект o_s и его модифицированную версию $o_{r_{copy}}$.

Аналогичные действия по созданию объектов-копий и их связыванию через отношение *истории* с исходными объектами, у которых были изменены идентифицирующие атрибуты, реализуются с помощью процедуры *NORMALIZE_ID*.

5. после выполнения всех действий, описанных в п. 4 результирующее множество содержит объекты, модифицированные точно таким же образом, как они были бы модифицированы при выполнении *OD.MT* над объектами *OD*-модели;

6. сохранить объекты из результирующего множества в метауровень. Это действие включает в себя также и создание новых объектов (п. 4), а также создание связей *истории*, *наследования* и *включения* между объектами (с использованием ODQL запросов).

После выполнения всех указанных действий будет получено взаимодействие $b \in B$, описывающее выполнение в точности тех же модификаций объектов, что изначально описывались алгоритмом f произвольной *OD*-модели. Доказательство теоремы служит основой для построения обобщённого алгоритма преобразования процедур изменения атрибутов объектов произвольной *OD*-модели в наборы взаимодействий.

4 Реализация инфраструктуры DIM

В более ранних работах, посвященных *DIM*, реализовывались отдельные компоненты СУБД, такие как компилятор запросов ODQL, однако, полноценной программной реализации СУБД, включающей как работу с данными, запросами, так и программирование в рамках *DIM* с учётом внесённых изменений в языки программирования реализовано не было.

Общей целью работы является разработка объединённой вычислительной инфраструктуры СУБД *DIM*, обладающей удобным пользовательским и пригодной к использованию в качестве СУБД, имеющей возможности, как хранения, так и обработки данных. Результаты проведённой работы могут быть использованы для программирования механизмов анализа данных и бизнес-логики приложений на основе разрабатываемой СУБД в проектах новых больших систем.

СУБД *DIM* была реализована поверх реляционной СУБД Oracle в виде дополнительных модулей, что позволило уйти от необходимости разработки физического уровня хранения данных, а также упростить сравнительный анализ доступности данных при их реляционной организации и при организации в виде *DIM* объектов.

DIM разделена на несколько взаимосвязанных компонентов, которые можно схематически представить следующим образом:

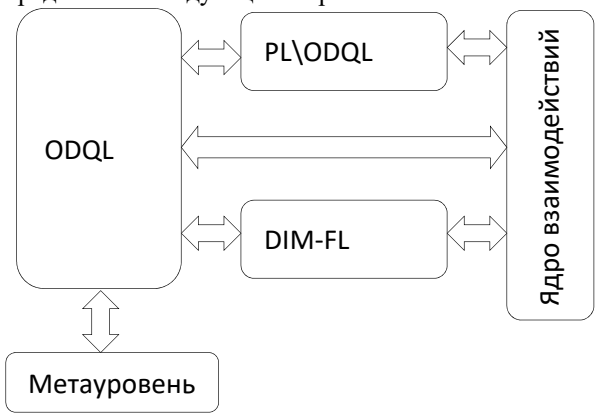


Рисунок 1 Компоненты *DIM*

4.1 Метауровень

Метауровень – уровень хранения данных, является заменой физического уровня данных. Представляет собой обертку над набором реляционных таблиц СУБД Oracle, используемых для хранения данных.

4.2 ODQL

Модуль ODQL, выполненный в виде Java-дополнения для «материнской» СУБД Oracle, служит для выполнения ODQL запросов выборки и изменения данных. При выполнении ODQL запроса происходит его преобразование до набора SQL запросов к исходным таблицам метауровня.

4.3 Реализация языков программирования

Языки программирования PL\ODQL и DIM-FL реализованы в виде соответствующих

интерпретаторов, встраиваемых в «материнскую» СУБД Oracle как Java-дополнения. При первичной загрузке исходных кодов модулей PL\ODQL или DIM-FL, хранимых в БД в виде объектов, происходит их преобразование в промежуточных код, и при дальнейших запусках функций из этих модулей повторная интерпретация исходного кода не требуется.

4.4 Ядро взаимодействий

Ядро выполнения взаимодействий является ключевой частью вычислительной инфраструктуры СУБД *DIM*, связывающей воедино интерпретаторы применяемых языков программирования, механизм выполнения ODQL запросов и метауровень доступа к данным. Далее схематично представлен механизм выполнения взаимодействий:

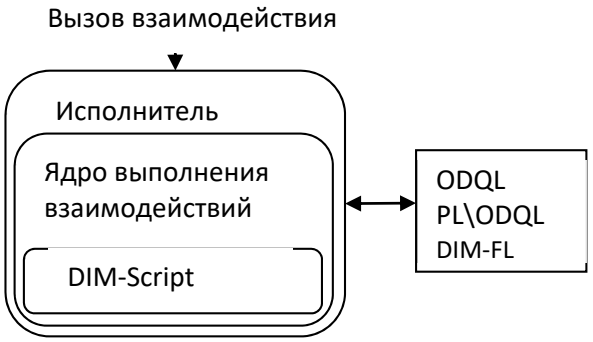


Рисунок 2 Механизм выполнения взаимодействий

Центральным модулем ядра выполнения взаимодействий является интерпретатор сценариев тела взаимодействия – «DIM-Script Interpteter», также далее называемый процессором взаимодействий. Интерпретатор строится на тех же принципах, что и интерпретаторы PL\ODQL и DIM-FL, однако реализует гораздо более простую логику выполнения ввиду очень ограниченного набора инструкций. С другой стороны, интерпретатор выступает и в роли контроллера выполнения взаимодействия, т.е. выполняет дополнительные операции по обеспечению передачи параметров в / из взаимодействия.

Выполнение взаимодействия процессором происходит следующим образом. Когда процессор получает вызов на выполнение взаимодействия, в первую очередь он загружает класс взаимодействия и определяет, является ли данное взаимодействие стартовым узлом графа взаимодействий либо нет. Если является, затем процессор загружает граф взаимодействий (загружает все содержащиеся в нём взаимодействия и воссоздаёт его структуру по загруженным связям между взаимодействиями). На следующем шаге контроль выполнения переходит к так называемой машине конечных состояний с множеством активных узлов (т.к. она может иметь несколько активных состояний одновременно: взаимодействия могут выполняться параллельно), где происходит выполнение всего графа взаимодействий.

Выполнение отдельного взаимодействия делегируется процессору взаимодействий. Сам процесс выполнения взаимодействия разделяется на несколько фаз:

1. Определение объектов, которые принимают участие во взаимодействии (объектов в ролях *Что*, *Откуда* и *Куда*) с помощью выполнения заданных во взаимодействии ODQL запросов либо передачи внешних параметров;
2. Проверка соблюдения предусловий, путём запуска специализированных проверочных PL\ODQL процедур (если при выполнении возникает исключение – условия не выполнены);
3. Выполнение вызовов модулей, входящих во взаимодействие путём делегирования выполнения вызовов PL\ODQL и DIM-FL интерпретаторов;
4. Сохранение изменённых объектов, выступавших в роли *Куда* в метауровень СУБД DIM.

4.5 Клиент DIM Developer

Для обеспечения возможности использования СУБД внешними приложениями был разработан программный интерфейс, предусматривающий возможность подключения удалённых клиентов, позволяющий работать как с данными, так и с их типами и управлять механизмами обработки данных. В основе механизма подключения – стандартный Java механизм подключения к БД JDBC, поскольку в качестве «материнской» СУБД используется реляционная СУБД. Таким образом, манипуляции с объектами *DIM* выполняются путём ODQL запросов, оборачиваемых в вызовы PL\SQL процедур «материнской» СУБД Oracle, являющихся, по сути, интерфейсом *DIM*.

Программирование прикладной логики СУБД DIM производится с помощью специально спроектированного клиентского приложения «DIM Developer», приближенного по функционалу к «Oracle SQL Developer». Клиент является разработанной на языке Java кроссплатформенной средой разработки прикладных модулей СУБД *DIM*. Среда позволяет выполнять ODQL запросы, создавать программные модули, а также наглядно проектировать взаимодействия и их объединения в цепочки и графы.

В результате исследовательской работы была построена развитая инфраструктура построения приложений, ориентированных на интенсивную работу с обширными объёмами различных взаимосвязанных данных. Были разработаны теоретические обоснования возможности использования аппарата *взаимодействий* для реализации алгоритмов произвольной *OD*-модели, перенесённых в инфраструктуру *DIM*. Проведенные практические исследования подтверждают применимость разработанной запросной технологии

ODQL и построенной поверх неё среды программирования для анализа и обработки данных.

Литература

[1] А.А. Барсегян, М.С. Куприянов, И.И. Холод, М.Д. Тесс, С.И. Елизаров. Анализ данных и процессов. СПб: БХВ-Петербург, 2009

[2] В.С. Рублев, Р.А. Юсупов. Концепции объектной динамической информационной модели DIM. Математика в Ярославском университете. Ярославль: ЯрГУ, 2006, С. 335-354.

[3] D.V. Antonov, V.S. Roublev, Effective interaction with the DIM DBMS. Proceedings of the Institute for System Programming Volume 27 (Issue 3). 2015 y. pp. 343-350.

[4] В.С. Рублев, А.Н. Петров, Язык PL\ODQL и множества с индексами. Ярославский педагогический вестник. Ярославль: ЯПГУ, 2012 г., 4: Т 3 (Естественные науки), С. 74-83.

[5] А. Н. Петров, Язык формул DIM-FL и его реализация в СУБД DIM. Молодая наука в классическом университете (Иваново, 21-25 апреля 2014г.). - Иваново. Издательство «Ивановский государственный университет», 2014. С. 44-45.

[6] В. С. Рублев. Теорема о статической полноте СУБД DIM. Проблемы теоретической кибернетики. Материалы XVII международной конференции. (Казань, 16-20 июня 2014г.). Казань: Отечество, 2014. С. 242-245.

[7] A.N. Petrov, V.S. Rublev Completeness of the Dynamics of the Attributes Values of Data in the Database DIM. Моделирование и анализ информационных систем. Ярославль: ЯрГУ, 2015 - 2. Vol.22.

Data analysis and processing using Dynamic Informational Model DB approaches

Alexey N. Petrov

Data analysis and processing mechanisms of Dynamic Informational Model (DIM) are considered in this paper. Generalized information about ways to organize and link data items in DIM are given in conjunction with specialized programing languages PL\ODQL, DIM-FL, DIM-Script and workflow-based approach to process data called interactions. The research extends and enhances PL\ODQL mechanisms for working with sets and indices. Also theoretical basis for completeness of object attributes dynamic is given and proven. To conclude the integrity of whole research and show its practical importance, DIM applications programming tool – «DIM Developer» and ways of the tool's usage with 3rd party apps are considered.

Сокращение числа виртуальных экспериментов с помощью оценки корреляций параметров взаимодействующих гипотез

© Е. А. Тарасов

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

Аннотация

В данной работе представлен подход, позволяющий исследователю сократить число виртуальных экспериментов, уменьшив количество наборов тестовых сценариев. Рассматриваемый подход основывается на вычислении корреляций между параметрами различных гипотез. Для решения данной задачи был выполнен обзор и сравнительный анализ существующих систем: Hephaestus, FCCE, ϒ-DB, реализующих схожий функционал. Далее, был произведен обзор алгоритмов отбора признаков, позволяющих уменьшить исследуемое пространство параметров и выявить взаимосвязь между ними. Были сформулированы функциональные требования к проектируемой системе. Рассмотрена практическая задача, которая может быть решена в рамках реализации данной платформы и описан ее частный случай, а именно, оценка корреляции параметров гипотез в астрономической задаче, которая будет использоваться в качестве тестового задания на этапе отладки системы.

Работа поддержана РФФИ (гранты 14-07-00548, 16-07-01028).

1 Введение

В современном мире исследования всё более зависимы от данных, которые становятся ключевым источником для получения новых знаний в той или иной области человеческой деятельности [4]. Такой подход получил название исследований с интенсивным использованием данных (ИИИД) [10] и развивается в соответствии с 4-й парадигмой научного развития [8]. Одним из ключевых элементов ИИИД является явное использование *гипотез* в определении *виртуального эксперимента* [3]. Гипотезам соответствует некоторая формальная спецификация свойств исследуемого явления, которая чаще всего имеет математическое представление. Сформулированные гипотезы нуждаются в тщательной проверке.

Труды XVIII Международной конференции DAMDID/RCDL'2016 «Аналитика и управление данными в областях с интенсивным использованием данных», Ершово, 11-14 октября 2016

В процессе выполнения виртуального эксперимента происходит манипулирование параметрами гипотез, т.е. набором переменных, которые в некоторых случаях могут быть коррелированы между собой, а также с параметрами других гипотез [10].

Так как число потенциальных гипотез в виртуальном эксперименте может быть огромным, а их взаимодействие нетривиальным, то в результате образуется пространство с большим числом виртуальных экспериментов, часть из которых плохо описывает наблюдения и нуждается в отсеивании ещё до выполнения эксперимента. Как следствие, исследователю необходимо средство, позволяющее заранее выявить и отсеять виртуальные эксперименты с прогнозируемо плохим результатом. В тоже время наличие сложных зависимостей в данных затрудняет их понимание исследователем и не позволяет делать это вручную [3]. Машинные средства оценки корреляции позволяют автоматически разделить виртуальные эксперименты на группы с заранее прогнозируемо хорошим и плохим результатом эксперимента [4].

Разрабатываемая архитектура платформы рассматривается в рамках более широкого проекта лаборатории, следовательно, будет интегрирована в него отдельным модулем [10].

Статья организована следующим образом. В разделе 2 приводится обзор существующих систем, позволяющих решать схожую задачу поиска корреляций и причинно-следственных связей. В разделе 3 приводится обзор алгоритмов отбора признаков. В разделе 4 формулируются требования к проектируемой системе. В разделе 5 формулируется тестовая задача поиска корреляции между двумя гипотезами. В разделе 6 формулируются дальнейшие шаги по развитию данной работы.

2 Обзор платформ для поиска корреляций над большими массивами данных

2.1 Hephaestus

В основе системы Hephaestus [3] лежит работа с виртуальными экспериментами над данными. Система помогает исследователю искать

корреляционные зависимости между большим числом переменных, предоставляет возможность сформировать гипотезы по наиболее перспективным связям, а затем с помощью тщательного тестирования перейти к причинно-следственной зависимости. Центральным блоком системы является SQL-подобный декларативный язык для описания виртуального эксперимента, с помощью которого возможно проектировать дизайн эксперимента, специфицировать основные гипотезы и их параметры, тестировать их, исполнять выбранные эксперименты и публиковать исследования. Так как применение только статистических методов и машинного обучения может приводить к ошибочным результатам в поиске причинно-следственных связей, то системы ориентирована на симбиоз между человеком и машиной. Используя корреляционные связи, которые были проверены экспертами в предметной области, Hephaestus собирает вероятностно причинные графы для спецификации семантики предметной области. Причинный граф поддерживает большое количество связей, обнаруженных в процессе выполнения виртуального эксперимента, а также позволяет исследовать последовательность выявления вероятностных причин и аномалий.

Hephaestus – это мета-система для исполнения виртуального эксперимента над существующими базами данных, которые могут уже выполнять некоторую аналитику локально. Она ориентирована на работу с очищенными и размеченными данными. Система состоит из следующих модулей:

- Получения набора данных. Авторы стремятся создать некую поисковую систему, которая принимает на вход строку, описывающую параметры гипотезы, возвращает ранжированный список потенциальных причинно-следственных зависимостей.
- Тестирования гипотез. Является основным рабочим модулем системы. Движок составляет запрос для оценки каждого возможного взаимодействия, разбивает образцы на контрольные блоки и высчитывает заданную метрику точности для каждого из них. После расчета статистики для блока, движок объединяет результаты, получая взвешенную оценку гипотезы.
- Ранжирования результатов. Гипотезы объединяются и сортируются по некоторой вероятностной оценке.

Вероятностно причинный граф – направленный ацикличный граф, содержащий коллекции причинно-следственных связей. Этот символьный язык позволяет исследователю интегрировать новые полученные знания для предметной области со своими ранее доступными знаниями в определенных областях науки.

2.2 FCCE

Платформа FCCE [16] разрабатывается для поиска корреляций в разнородных наборах данных, охватывающих большие временные диапазоны. Она ориентирована на работу с минимальными задержками и доступом к не пред-обработанным исходным данным.

Данная система рассматривалась на примере 2-х задач из области безопасности: обнаружения доменных имен потенциальных сетей зараженных рабочих станций и пост-инцидентное расследование проникновения.

Для минимизации задержек обработки данных решено было отказаться от использования традиционных реляционных баз данных для хранения информации и перейти на NoSQL.

Ключевым компонентом модели данных является концепция признаков. Признаки определяют связь между парой ключ-значение, каждый элемент из которых может содержать несколько атрибутов. FCCE представляет упрощенную реляционную модель данных для пользователя, где каждая таблица хранит один тип признаков. Каждая строка идентифицируема ключом и может содержать несколько атрибутов.

FCCE обеспечивает API для хранения, получения и вычисления корреляции над признаками. Разработчики предлагают оригинальный подход интеграции модуля оценки критерия корреляции в движок исполнения запросов над хранилищем, позволяющий ускорить время ответа на запрос и снизить накладные расходы на вычисление и ввод-вывод. FCCE использует два отличительных механизма для поддержки эффективности операций нахождения корреляций между признаками: канал запросов и модификатор запросов. Канал – механизм, позволяющий передавать признаки, извлеченные из одного запроса в другой запрос в качестве входных данных, т.е. последовательно можно объединять несколько GET функций, тем самым создавай пересечения нескольких признаков. Модификатор – над GET запросом предполагает использование широкого набора опций для более тонкого контроля его поведения.

Архитектура платформы состоит из следующих модулей:

- Извлечение. Для каждого источника эксперты в области определяют метод извлечения признаков из сырых данных.
- Агрегация. Данные собираются из различных локальных экстракторов в так называемые коллекторы, которые выполняют функцию дедупликации, отказоустойчивости, балансировки нагрузки.
- Хранение. Централизованное хранилище, над которым выполняются запросы к признакам.
- Получение. Модуль обеспечивает интерфейс запросов над хранилищем признаков. Доступ к данным организован с помощью 3 компонент. Первый состоит из Сервиса регистрации, осуществляет поиск корневого коллектора, и Протокола запросов, посылает запросы к соответствующему хранилищу,

используя тип признаков и ключи в качестве предикатов запросов. Второй, используя специальный протокол, может подписаться на определенный экстрактор или коллектор, так что при появлении интересующей пары ключ-значение они сразу попадут в него. Третий реализует интерфейс поиска корреляции признаков и позволяет настроить различные функции корреляции для получения знаний из различных типов признаков.

Более подробно механизм поиска корреляции признаков авторы статьи собираются раскрыть в будущих работах. В перспективе они так же планируют перенести функционал поиска корреляций с уровня доступа к данным на уровень хранилища признаков для уменьшения задержки при обработке сложных запросов.

2.3 Ƴ-DB

Разработки системы Ƴ-DB [4, 5] ведутся с целью поддержки процесса проведения научных исследований, обеспечивая возможность управления гипотезами и их анализа. Предиктивная аналитика строится над вероятностной базой данных.

В работе делается упор на управление параметрами гипотез. Ключевые особенности такого подхода и их отличия от управления экспериментальными данными заключаются в следующем:

- Работа ведется не со всеми данными, полученными в результате эксперимента, а только лишь с некоторым отобранным подмножеством. Тем самым уменьшается объем, но увеличивается структурированность данных.
- Если при работе с обычными данными модель доступа к ним ориентирована на работу с измерениями (денормализованный вид), то модель хранения параметров гипотез определяется из её структуры, т.е. происходит нормализация по факторам неопределенности.
- К неопределенностям на уровне данных, источниками которой являются их неполнота и несогласованность, так же добавляется неопределенность, порожденная существованием множества конкурирующих гипотез.

В качестве примера авторами был разобран сценарий расчета физиологических гипотез, а именно тестирование трёх различных теоретических моделей насыщенности гемоглобина кислородом.

Первый этап работы с гипотезами – это их кодирование. Для вычисления предсказаний гипотезы используют асимметричные функции, которые выполняют оценку над входными переменными (параметры) для вычисления значений выходных переменных (предсказаний). Техника кодирования гипотез базируется на наличие структуры гипотезы в машиночитаемом формате W3C MathML. Платформа Ƴ-DB имеет XML адаптер для извлечения моделей зашифрованных в формате MathML и вывода причинных зависимостей.

Ключевым компонентом архитектуры платформы является канал синтеза, который представляет собой последовательный процесс обработки данных. На вход поступает структура гипотез и их данные. Из структуры извлекается функциональные зависимости. Данные помещаются в большую таблицу, содержащую все переменные как реляционные атрибуты в таблице. Затем включается компонент синтеза и трансформирует данные из большой таблицы в вероятностную базу данных, где каждая гипотеза декомпозируется в таблицы претендентов. Авторами был предложен алгоритм трансформации каждой гипотезы в вероятностную таблицу. Базовый принцип проектирования неопределенного моделирования состоит в том, чтобы определить только одну случайную переменную для каждого действительного фактора неопределенности (u-фактор). Модель гипотезы сама по себе это теоретический u-фактор, чья неопределенность исходит из множества моделей, ориентированных на объяснение того же явления. Множество испытаний каждой гипотезы нацеленных на один и тот же феномен порождает множество эмпирических u-факторов. Для поддержки тестирования гипотезы вероятностное распределение феномена должно учитывать оба вида u-фактора.

Предиктивная аналитика выполняется над вероятностной базой. Ƴ-DB не предлагает каких либо новых инструментов для тестирования гипотез. Насколько можно понять, это статические методы на основе Байеса.

Прототип системы разработан как Web-приложение, написанное на Java, с компонентами канала, реализованными на стороне сервера поверх MayBMS. Где MayBMS – это расширение PostgreSQL. Как отмечают авторы управление данными гипотез является перспективным новым полем исследовательской деятельности, позволяющим получить больше пользы из эксперементальных данных, открытых исследовательскими лабораториями. В планы дальнейшего развития входит улучшить: статистические способности и масштабируемость системы для тестирования выборок большого объема.

2.4 Отличия от существующих подходов

В рамка данного подхода исследователь работает с уже существующими гипотезами, моделирующими свойство какого-либо явления в природе, экономике, бизнесе и т.д. Гипотеза является ключевым элементом рассматриваемого метода. Все параметры гипотез являются ценными и несущими информацию и поэтому от них нельзя избавляться.

Из-за присутствия априорных знаний, накопленных в виде гипотез, выбор параметров не

является полностью черным ящиком – в отличие от методов машинного обучения. Исследователю также заранее известна некоторая часть взаимосвязей между гипотезами, т.е. какие из них зависимы друг от друга.

Для части гипотез могут быть доступны локальные наблюдения, соответствующие их параметрам и выступающих в качестве ограничений на совокупность этих гипотез. Кроме наблюдений может быть доступна некоторое теоретическое распределение параметров гипотез.

Гипотезы могут быть сформулированы как набор правил, система математических уравнений и пр. Разрабатываемая в рамках данного подхода система должна уметь работать с разнообразными входными данными. При добавлении новых значений параметров гипотез или включений новых гипотез в существующий набор система должна обновлять набор сокращенных экспериментов.

Предлагаемый метод работает с симуляциями и наблюдениями. Сопоставляя экспериментальные и фактические данные, мы восстанавливаем полную модель по частично доступной.

Рассматриваемый подход нацелен на исключение заведомо «плохих» экспериментов, т.е. тех, которые производят симуляции с большой ошибкой. Установление причинно-следственных связей не является целью данной работы. Возможность поддержки данного механизма планируется в дальнейшем.

Таким образом сокращение числа экспериментов достигается за счет:

- Поиска корреляций – это позволяет объединить признаки исследуемого явления в некоторые группы, оказывающие влияние на него в некоторой совокупности.
- Анализ этих признаков – необходимо подобрать набор значений параметров в рамках выделенной группы, которые с определенными показателями точности описывали бы фактические данные наблюдений.
- Ранжирование гипотез по степени точности виртуального эксперимента. Это поможет исследователю обратить внимание на наиболее вероятные гипотезы без необходимости полного перебора гипотез.

3 Методы отбора признаков

Для уменьшения пространства параметров гипотез и виртуальных экспериментов используются методы отбора признаков [13]. Они позволяют увеличить скорость обработки данных и получения результат, не снижая показатели точности [18], путем выделения только тех информативных признаков, которые требуются для выполнения виртуального эксперимента.

Выделение набора признаков позволяет упростить понимание модели исследователем и, следовательно, использовать их в качестве входных данных для широко известных алгоритмов машинного обучения [6]. Так же данные методы позволяют уменьшить шум в данных и выявить взаимодействие между параметрами.

Методы отбора признаков возможно классифицировать следующим образом: Фильтры, Обертки, Встроенные [1, 13].

Фильтры. Опираются на общие характеристики обучающих данных и осуществляют процесс выборки признаков в качестве шага предварительной обработки независимо от индукционного алгоритма. Обладают низкой стоимостью вычислений. Фильтры используются в кластеризации для построения начального приближения. Не предназначены для выявления сложных связей между признаками, т.к. обладают низкой чувствительностью.

К таким методам можно отнести: **CFS** [6] – где выбор признаков на основе корреляций. Является простым многофакторным фильтрующим алгоритмом, который раскладывает подмножество признаков согласно эвристической функции оценки, основанной на корреляции. **INTERACT** [20] – двух этапный алгоритм, основанный на симметричной неопределенности и согласованности. **ReliefF** [11] – который является расширением алгоритма Relief, и работает путем случайной выборки экземпляра из данных, а затем находит его ближайшего соседа из того же или противоположного класса. **mRMR** [14] – выбирает признаки, которые имеют самое высокое значение информативности с целевым классом и обладающие минимальной избыточностью. Его разновидностью является **M_d фильтр** [17] – который использует меру монотонной зависимости для оценки информативности.

Обертки. Включают оптимизацию предиктора как часть процесса выбора. Позволяют выявлять зависимости признаков. Качество выборки зависит от индукционного алгоритма. Основным недостатком является вычислительная нагрузка, которая исходит от вызова алгоритма индукции для оценки каждого подмножества интересующих параметров.

К ним можно отнести: **WrapperSubsetEval** [19] – вычисляет наборы признаков с использованием схемы обучения. Для оценки точности схемы обучения для набора признаков используется перекрестная проверка. В качестве схемы обучения могут использоваться SVM и C4.5.

Встроенные. Выполняют функции выборки в процессе обучения. Как правило специфичны для алгоритмов машинного обучения. Применимость метода всегда зависит от типа решаемой задачи. Позволяют выявлять зависимости признаков. Обладают хорошей скоростью работы.

К ним относятся: **SVM-RFE** [15] – метод осуществляет выбор признаков итеративным обучением SVM классификатора с текущим набором признаков и удаляет наименее важный признак, указанный SVM. Существуют две версии этого метода: с линейным и нелинейным ядром. **FS-P** [12]

– основанный на перцептроне. Идея метода заключается в обучении перцептрона в контексте контролируемого обучения. Веса взаимосвязей используются как индикатор того, какие признаки могут быть наиболее информативными.

Другие методы. Также к методам, позволяющим снизить размерность данных и оценить зависимость параметров, можно отнести следующие техники. Анализ главных компонент (**PCA**) [7], которая включает в себя преобразование ряда коррелируемых переменных в меньшее число не коррелируемых. Анализ независимых компонент (**ICA**) [9], позволяющий не только декоррелировать параметры, но также уменьшает статистические зависимости более высокого порядка. Канонический корреляционный анализ (**CCA**) [7], устанавливающий соотношения линейных связей между двумя многомерными переменными. Неотрицательная факторизация (**NMF**) [9], позволяющая накладывать дополнительные ограничения на главные компоненты.

Применимость того или иного метода в разрабатываемой системе будет оценена на этапе отладки в рамках решения тестового сценария, описанного ниже.

4 Формализация требований

Разрабатываемая платформа должна удовлетворять следующим требованиям:

- Система должна быть модульной и функционально расширяемой.
- Должна поддерживать связность с другими компонентами глобальной системы, в рамках которой она реализуется.
- В качестве модуля хранения данных должна использоваться платформа, ориентированная на работу с большими объемами, а также поддерживающая современные средства аналитики.
- Ключевым компонентом системы является модуль по поиску корреляций параметров гипотез. В качестве используемых методов предлагается использовать различные подходы: байесовский, частотный, методы машинного обучения и сравнить полученными ими результаты между собой.
- Системы должна иметь возможность работать с уже сформулированными гипотезами. Гипотезы должны хранится в базе данных на ряду с экспериментальными данными и результатами проведения виртуального эксперимента.
- В зависимости от практической задачи методы машинного обучения могут отличаться от описанных в предыдущем разделе.

5 Сценарий для тестирования

В качестве сценария для тестирования будет рассмотрена частная задача Безансонской Модели Галактики [2] о нахождении корреляции параметров между двумя независимыми гипотезами, а именно Star Formation Rate и Initial Mass Function. Данные гипотезы описывают процесс зарождения звезды. Известно, что параметр γ в SFR коррелирует с моделями IMF. Как отмечают авторы:

“Однако, мы хотим подчеркнуть, что параметр γ коррелирует со значениями других параметров, используемых в модели, и особенно склонами (α) в IMF и возраста диска.” [2]

В настоящее время авторы фиксируют параметры остальных гипотез и изучают влияние α и γ на поведение модели вручную. Соответственно, в данном случае наша система значительно бы облегчила работу исследователей.

Гипотеза SFR представляет общую массу звезд, зарожденных в определенной области Галактики за некоторый интервал времени. В версии БМГ от 2014 года [2] функция представляется авторами в следующем виде:

$$SFR(i) = \exp(\gamma \times x_c(i)) \times d$$

Где γ – исследуемый параметр, x_c – возраст в i-ом интервале, d – размер возрастного интервала.

Гипотеза IMF представляет функцию распределения массы определенной популяции звезд. В общем виде может быть представлена:

$$\phi(m) = m^{-\alpha}$$

Где m – масса, α – параметр, характеризующий склон функции. Так как функция представлена на трех интервалах, то, таким образом, относительная масса внутри каждого интервала может рассчитываться как:

$$K_i \int_{m_i}^{m_{i+1}} m\phi(m)dm = MInt_i$$

Где K_i – коэффициент непрерывности, i – рассматриваемый интервал.

В данном примере представлены две гипотезы имеющие математическое представление и предположение о взаимной зависимости их параметров α и γ.

На первом шаге оценивается корреляция параметров данных двух гипотез. Значения оценки, полученные различными методами поиска корреляций, должны иметь сопоставимые между собой значения.

На втором шаге выполняется анализ данных признаков. Выполняя виртуальные эксперименты над ними и сравнивая результаты с реальными наблюдениями, накапливается информация об исследуемой модели.

Полученные значения дают возможность ранжировать гипотезы и выбрать наиболее вероятные, тем самым упрощая процесс выбора значения параметров гипотез модели. Результатом данной процедуры является набор значений параметров α и γ, при этом ожидается что описанные

ранее авторами параметры [2] будут включены в этот набор.

Заключение

В данной работе описан подход, позволяющий уменьшить пространство возможных гипотез в виртуальном эксперименте. Его идея базируется на поиске и оценке корреляций между параметрами выбранных гипотез. Дан обзор существующих и проектируемых систем, которые в той или иной мере реализуют функционал поиска корреляций над большими массивами данных. Представлены некоторые алгоритмы оценки взаимосвязи параметров, с учетом специфики работы с астрономическими данными [9]. Сформулированы требования, предъявляемые к проектируемой архитектуре.

В качестве дальнейших шагов развития подхода планируется реализация системы с учетом описанных ранее требований. В качестве практической задачи будут исследованы корреляции всех параметров гипотез Безансонской Модели Галактики [2], взаимосвязь двух гипотез которой рассматривалась в рамках тестовой задачи.

Благодарность

Автор статьи выражает благодарность Д. Ю. Ковалеву за предоставленную идею.

Литература

[1] Veronica Bolon-Canedo, Noelia Sanchez-Marono and Amparo Alonso-Betanzos. A review of feature selection methods on synthetic data. Knowledge and Information Systems, 34(3), p. 483-519, 2013

[2] Maria A. Czekaj, Annie C. Robin, Francesca Figueras and Xavier Luri. Galaxy evolution: A new version of the Besancon Galaxy Model constrained with Tycho data. Barcelona: Universitet de Barcelona, PhD Thesis, 2012

[3] Jennie Duggan and Michael Brodie. Hephaestus: Data Reuse for Accelerating Scientific Discovery. In Proceedings of 7th Biennial Conference on Innovative Data Systems Research (CIDR'15), Asilomar, California, USA, 2015

[4] Bernardo Goncalves and Fabio Porto. Managing large-scale scientific hypotheses as uncertain and probabilistic data with support for predictive analytics. IEEE Computing in Science and Engineering, 17(5), p. 35-43, 2015

[5] Bernardo Goncalves, Frederico C. Silva and Fabio Porto. ϒ-DB: A system for data-driven hypothesis management and analytics, 2014. http://arxiv.org/abs/1411.7419

[6] Mark A. Hall. Correlation-based Feature Selection for Machine Learning. The University of Waikato, Hamilton, New Zeland, PhD Thesis, 1999

[7] David R. Hardoon, Sandor Szedmak and John Shawe-Taylor. Canonical correlation analysis: an overview with application to learning methods. Neural Computation, 16(12), p. 2639-2664, 2004

[8] Tony Hey, Stewart Tansley and Kristin Tolle. The Fourth paradigm: Data-intensive scientific discovery. Redmond, Microsoft Research, 2009

[9] Zeljko Ivezic, Andrew J. Connolly, Jacob T. VanderPlas and Alexander Gray. Statistics, data mining, and machine learning in astronomy: A practical Python guide for the analysis of survey data. Princeton University Press, 2014

[10] Leonid Kalinichenko, Dmitry Kovalev, Dana Kovaleva and Oleg Malkov. Methods and tools for hypothesis-driven research support: a survey. Informatica and Appications, 9(1), p. 28-54, 2015

[11] Igor Kononenko. Estimating attributes: analysis and extensions of RELIEF. In Proceedings of the European conference on machine learning (ECML'94), Catania, Italy, p. 171–182, 1994

[12] Manuel Mejia-Lavalle, Enrique Sucar and Gustavo Arroyo. Feature selection with a perceptron neural net. In Proceedings of the international workshop on feature selection for data mining: Interfacing Machine Learning and Statistics, p. 131–135, 2006

[13] Luis C. Molina, Lluis Belanche and Angela Nebot. Feature Selection Algorithms: A Survey and Experimental Evaluation. Data Mining, 2002. ICDM 2003. In Proceedings of 2002 IEEE International Conference on Data Mining, p. 306-313, 2002

[14] Hanchuan Peng, Fuhui Long and Chris Ding. Feature selection based on mutual information: criteria of max-dependency, max-relevance, and min-redundancy. IEEE Transactions on Pattern Analysis and Machine Intelligence, 27(8), p. 1226–1238, 2005

[15] Alain Rakotomamonjy. Variable selection using SVM-based criteria. The Journal of Machine Learning Research, 3, p. 1357-1370, 2003

[16] Douglas Schales, Xin Hu, Jiyong Jang, Reiner Sailer, Marc Stoecklin and Ting Wang. FCCE: Highly Scalable Distributed Feature Collection and Correlation Engine for Low Latency Big Data Analytics, In Proceeding of 2015 IEEE 31st International Conference on Data Engineering, p. 1316–1327, Seoul, IBM Research Report, 2014

[17] Sohan Seth and Jose C. Prıncipe. Variable selection: A statistical dependence perspective. In Proceedings of the international conference of machine learning and applications (ICMLA'10), p. 931–936, 2010

[18] Nigel Williams, Sebastian Zander and Grenville Armitage. A Preliminary Performance Comparison of Five Machine Learning Algorithms for Practical IP Traffic Flow Classification. ACM SIGCOMM Computer Communication Review, 36(5), p. 5-16, 2006

[19] Ian H. Witten and Eibe Frank. Data mining: practical machine learning tools and techniques. Morgan Kaufmann Publishers, San Francisco, 2005

[20] Zheng Zhao and Huan Liu. Searching for interacting features. In Proceedings of the international joint conference on artificial intelligence (IJCAI'07), p 1156–1161, Hyderabad, India, 2007

Reducing the number of virtual experiments by estimating the correlation parameters of interacting hypotheses

Evgeny Tarasov

This paper presents the approach that helps to researcher to reduce the number of virtual experiments through decrease the count of tested hypotheses. This approach is based on correlation search between parameters of different hypotheses. A review and analysis of modern platforms with similar functionality is done. Methods for reducing the number of virtual experiments are surveyed, including the features selection algorithm, which allows to reduce investigated parameters space and identify the interaction between them. Next, functional requirements of designed system are formulated. We consider the practical problem which can be solved in the framework of this system and consider its particular case – assessment of the correlation parameters in astronomical hypotheses problem, which will be used as the test task during system debugging.

Открытый симпозиум:
Интенсивное использование данных в здоровьесбережении

Open Workshop:
Data-Intensive Healthcare

Концептуальные основы и архитектура интернет-системы персонализированной поддержки здоровьесбережения на основе интенсивного анализа данных

© В. Н. Крутько © А. И. Молодченков

Федеральный исследовательский центр «Информатика и управление»
Российской академии наук,
Москва

krutkovn@mail.ru aim@isa.ru

Аннотация

В работе охарактеризована проблема здоровья населения России и пути ее решения, сформулированы цели и описаны концептуальные основы системы здоровьесбережения, охарактеризована структура пространства управления здоровьем и общая архитектура интернет-системы персонализированной поддержки здоровьесбережения, а также входящие в систему модули и сервисы. Основными функциями системы являются: сбор информации из различных источников, включая интернет-пространство; интеллектуальная обработка медицинских данных и текстов; анализ эффективности технологий здоровьесбережения; оценка проблемных зон в здоровье конкретного человека; персонализированная оптимизация технологий его здоровьесбережения; помощь в их применении и мониторинг их эффективности.

Исследование выполнено при финансовой поддержке Минобрнауки РФ в рамках проекта № 14.607.21.0123.

1 Введение

По имеющимся оценкам Всемирного Банка и прогнозам Минэкономразвития [5] медико-демографическая ситуация в России является одним из главных препятствий эффективному социально-экономическому развитию страны в настоящем и будущем. Несмотря на наблюдающиеся в последние годы позитивные сдвиги, отставание России по показателям здоровья не только от развитых, но и от и многих развивающихся стран мира в настоящее время очень велико и продолжает возрастать. По данным ВОЗ [8], в 2013 г. ожидаемая продолжительность жизни (ОПЖ – главный показатель здоровья нации) была, по сравнению со странами-рекордсменами, ниже на 19 лет у мужчин и на 12 лет у женщин. Среди 194 стран – членов ВОЗ Россия занимала 144-е место по ОПЖ мужчин и 106-е по ОПЖ женщин. Это значительно хуже показателей 1990 г. По ожидаемой продолжительности здоровой жизни (ОПЗЖ) мужчин в 2013 г. Россия разделила места с 144-го по 146-е с Руандой и Сенегалом. Отставание по ОПЗЖ от страны-лидера по этому показателю – Сингапура – составило 20 лет для российских мужчин и 12 лет для женщин. Сравнив эти значения с уровнями отставания по ОПЖ, нетрудно заметить, что при значительно меньшей, чем в наиболее благополучных странах, продолжительности жизни российские женщины несут равное бремя болезней (исчисляемое как ожидаемое число лет, проведенных в состоянии болезни и/или инвалидности), а российские мужчины – даже большее. По смертности мужчин в возрастной группе 60-64 года Россия в 2013 г. заняла первое место среди всех 194 стран – членов ВОЗ! Таким образом, россияне живут не только меньше, но и существенно хуже, чем жители развитых стран.

Проблемы со здоровьем обусловливают огромные потери, которые несет как государство в целом, так и каждый отдельно взятый россиянин. Bloom, Canning & Sevilla [7] показали, что здоровье населения является критической составляющей экономического роста. Вклад этого фактора в совокупный объем национального производства более значим, чем эффекты других параметров человеческого капитала - уровня образования и профессионального опыта работающих. Увеличение ОПЖ на 1 год дает прирост ВВП на 4%. Поэтому подтягивание России по величине ОПЖ до уровня развитых стран может обусловить прирост ВВП страны на величину порядка 50%.

Эффективным подходом к решению вышеназванных проблем является использование высоких информационных технологий,

Труды XVIII Международной конференции DAMDID/RCDL'2016 «Аналитика и управление данными в областях с интенсивным использованием данных», Ершово, 11-14 октября 2016

позволяющих помочь каждому жителю России осознать важность здоровья, получить мотивацию к его сбережению, получить персонализированные рекомендации и в результате качественно решить свои проблемы со здоровьем.

На сегодняшний момент разработано достаточно много систем и мобильных приложений по поддержке здорового образа жизни и слежения за своим здоровьем. Однако большинство из них не дают научно-обоснованные персонализированные рекомендации по здоровьесбережению. Отсюда возникает задача разработки такой технологии по поддержке здоровьесбережения, которая бы позволяла давать персонализированные рекомендации по поддержке здоровья пользователей. При этом данная технология должна предоставлять возможности дополнения функционала к уже существующим на рынке решениям.

В настоящей работе представлены концептуальные основы и архитектура интернет-системы персонализированного здоровьесбережения, базирующейся на интенсивном анализе данных.

Работа выполнена при финансовой поддержке Министерства образования и науки РФ, соглашение №14.607.21.0123, проект «Разработка интернет-технологии для персонализированной поддержки здоровьесбережения».

2 Концептуальные основы системы здоровьесбережения

Давно известно, что основными средствами формирования здоровья являются профилактика заболеваний и здоровый образ жизни (ЗОЖ). Приоритет профилактики и ЗОЖ постулирован в национальном проекте «Здоровье», в Государственной программе РФ «Развитие здравоохранения» [2], в т. наз. «майских указах» Президента РФ (2012 г.) и подтвержден в мировом масштабе ВОЗ («Глобальная стратегия по питанию, физической активности и здоровью») [1] и ООН (Резолюция круглого стола «Изменяя мир: укрепление здорового образа жизни и контроля за неинфекционными заболеваниями» Генеральной ассамблеи ООН, Нью-Йорк, 2015 г.,) [6]. Большое внимание проблемам здоровья уделяет мировая наука – за последние пять лет насчитывается более 900 000 публикаций, затрагивающих проблему «public health».

Реализуемый нами проект по сути направлен на преобразование ЗОЖ в ЗОЖ-HiTech.

Глобальной цель проекта является создание интернет-системы персонализированной поддержки здоровьесбережения (далее «Системы Здоровьесбережения» или «СЗ»), помогающей обеспечить «Достижение максимальной продолжительности максимально активной, созидательной, полноценной и эмоционально позитивной жизни каждого человека».

Для достижения этой цели необходимо сконцентрировать усилия на двух задачах:

1) увеличение текущего уровня здоровья (уровня психической и физической работоспособности и эмоциональной позитивности);

2) увеличение периода активной жизни человека.

Поставленная задача является глобальной и структурно сложной, поэтому на первом этапе проектирования системы был разработан ее концептуальный базис, обеспечивающий эффективное достижение поставленной цели.

Данный базис представляет собой нижеследующую систему принципов и требований, которые отражают основные важные характеристики объекта управления – «здоровья человека» [3, 4] и должны быть учтены и обеспечены конструкцией СЗ.

2.1 Концептуальный базис системы здоровьесбережения

Принцип полноты.

Принцип полноты понимается как стремление к максимально полному представлению в Системе Здоровьесбережения всех наиболее значимых процессов, определяющих здоровье человека. Это представление осуществляется в N-мерном «Пространстве управления персональным здоровьем» или проще «Пространстве Здоровьесбережения», где оси данного пространства имеет следующий смысл: 1) показатели здоровья; 2) факторы, определяющие здоровье; 3) источники данных о здоровье и определяющих его факторах. 4) методы анализа данных здоровьесбережения; 5) методы управления здоровьем; 6) проблемы здоровья и болезней; 7) витальный цикл человека; 8) все неоднородное население России (и в перспективе – всего мира); 9) сервисы Системы Здоровьесбережения.

По каждой координате необходимо стремление к наиболее полному набору возможных данных и методов. Принцип полноты обеспечивается контрактами с клиниками (полнота методов диагностики и лечения) и с владельцами методов анализа и технологий управления здоровьем. Необходимо стремиться к максимуму информации о пользователях во всех возможных формах, сочетаемых с методами извлечения информации (data mining), распознавания и анализа, в том числе интеллектуального (cognitive). Такие полные данные являются ценными сами по себе, реализуют функцию «биобанка» и могут при коммерциализации проекта продаваться заинтересованным организациям и фирмам.

Принципы системности.

Требование базирования на Био-психо-социо-духовной концепции здоровья – охвате всех важнейших для здоровья сфер жизнедеятельности человека, привлечении мотивационных, эмоциональных, интеллектуальных, социальных и духовных ресурсов, то есть, организация

поддерживающей ЗОЖ среды во всех сферах функционирования человека.

Проблема здоровьесбережения рассматривается в целостной системе «среда-организм».

Представление организма человека как органически целостной иерархической системы, где можно выделить разные уровни воздействия и разные жизненно важные функциональные подсистемы.

Определение связей с надсистемой, в которую погружена создаваемая система здоровьесбережения, а также определение подсистем системы и связей между ними – структуры системы.

Принцип открытого эволюционного развития

Масштабность системы, постоянное появление новых знаний о здоровье и методов его коррекции требует гибкой, открытой, модульной конструкции системы, позволяющей легко ее развивать качественно и количественно, дополнять базы новой информацией, а систему новыми алгоритмами и методами, а также подключать к системе внешние сервисы и др. системы.

Принцип оптимального баланса между консерватизмом и революционностью.

Оптимальный баланс между консервативным подходом, предполагающим учет в системе всего ценного, что наработало человечество в области здоровьесбережения (медицинские протоколы в клинике, руководства по профилактической медицине ВОЗ и ведущих стран - здесь наша новизна в обобщении, интеграции, информационной и мотивационной упаковке имеющихся знаний, а также в алгоритмах использования этих знаний для персонализированной оптимизации здоровьесбережения) и революционным подходом - использованием новых прорывных технологий (ОМИКсные технологии, инфо-когнитивные методы и др.). Здесь интересно упомянуть результаты международного исследования, показывающие, что имеется оптимальное соотношение между консервативным и революционным элементами, равное 2/1, которое обеспечивает наиболее эффективное социально-экономическое развития разных стран.

Принцип иерархической эффективности.

Принцип полноты предполагает бесконечное расширение и совершенствование системы, что создает опасность утонуть, затеряться в бесконечном разнообразии проблем здоровья и методов его улучшения. Поэтому принципиально важно с самого начала ставить задачи экспертного или аналитического выбора наиболее важных и значимых для здоровья элементов по каждой из координат N-мерного «Пространства Здоровьесбережения». Необходимо строить иерархические списки этих элементов по степени их значимости для здоровья.

Такой подход подкрепляется известным принципом Парето - 80% эффектов деятельности системы определяют 20% наиболее важных элементов системы.

Мыслить глобально, а действовать локально – с одной стороны, постоянно обеспечивать принцип полноты в подходе и общей организации структуры системы, а с другой стороны – обеспечивать модульность структуры системы и последовательность конкретных законченных шагов по разработке наиболее интересных модулей – последовательность законченных версий СЗ.

Принцип управления целями – метод управления, лежащий в основе технологии эффективного массового здоровьесбережения, заключается в управлении информационной средой обитания человека, в которую человек погружен: формулировка целей и описание желаемых эффектов, разработка эффективных методов и технологий достижения целей, информирование людей об этих целях, методах и технологиях, информационная помощь в их персонализированном практическом применении.

Принцип персонифицированной оптимизации – оптимизация технологий здоровьесбережения для данного конкретного человека с учетом его психо-физиологических характеристик, уровня и образа жизни, характеристик социо-природной среды его обитания – его интегральный паспорт здоровья.

Принцип стандартизации – должны использоваться, по аналогии с медицинскими стандартами лечения, имеющиеся стандарты профилактической медицины, а также разрабатываться новые стандарты для новых персонализированных профилактических программ и новых средств и методов информатики, используемых для оценки эффективности средств и методов оздоровления, для персонализированной оптимизации и поддержки применения оздоровительных программ, для мониторинга достигнутых результатов и полученных эффектов.

Критерии качества системы показателей здоровья: полнота; специфичность; структурированность; оптимальность; измеримость и ясность; управляемость (эластичность); информационная насыщенность; значимость; интерпретируемость (прогностическая сила). В совокупности показатели должны позволять вычислять на их основе наиболее важные, но иногда может быть прямо не измеримые характеристики здоровья.

Требование применения современных методов оценки важности показателей и факторов, определяющих здоровье: методов интеллектуального анализа данных из социальных сетей, методов лингвистического анализа текстов (морфологического, синтаксического и семантического анализа); методов машинного обучения.

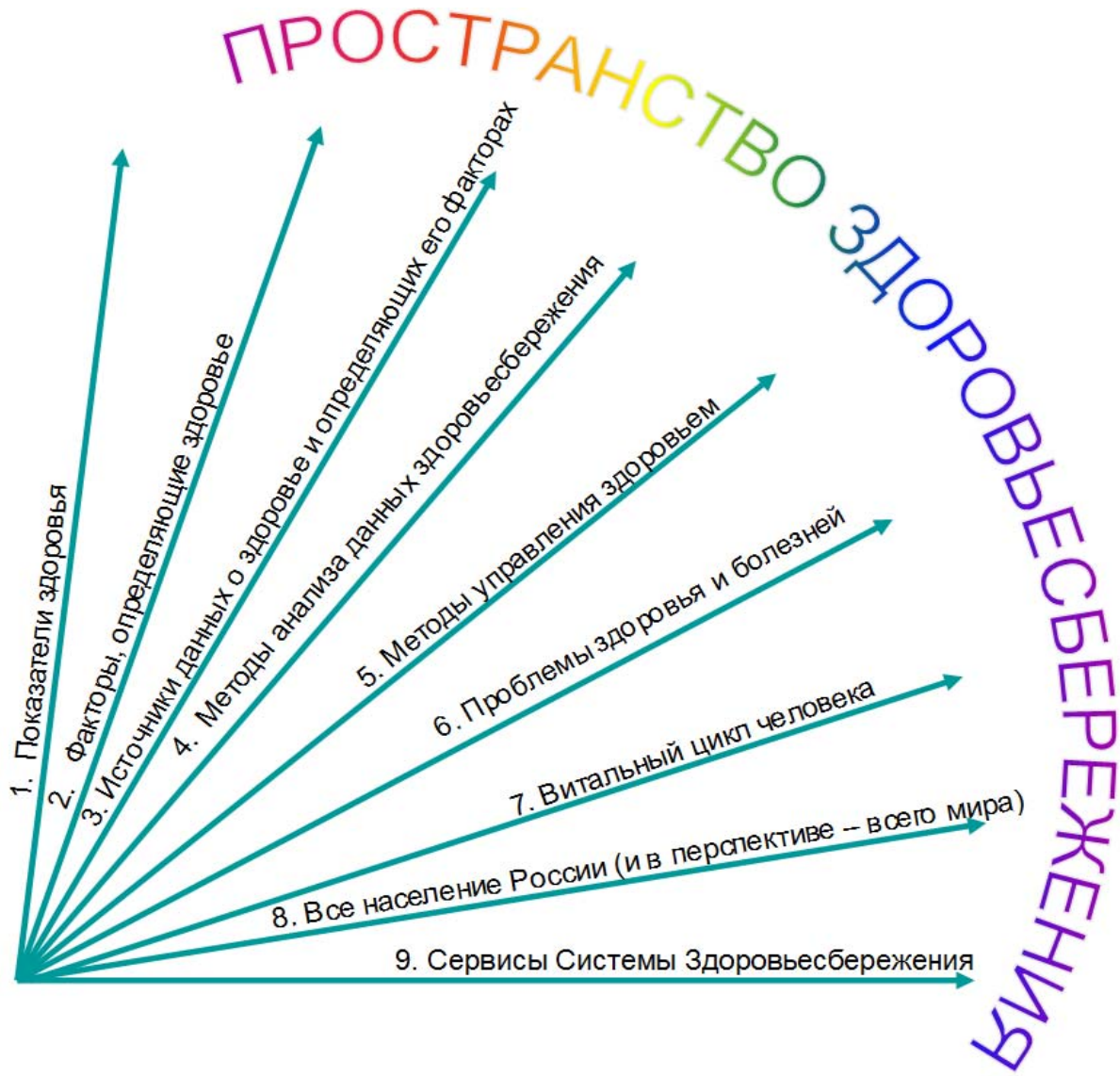


Рисунок 1 Пространство здоровьесбережения

Принцип мотивации и психологической поддержки здоровьесбережения - должна быть обеспечена персонализированная мотивация населения по применению ЗОЖ с учетом личностных характеристик человека: за счет надежности и обоснованности информации об эффективности ЗОЖ; за счет применения современных методов рекламы и убеждения через средства массовой информации и интернет; за счет системы стимулов быть здоровым, организуемой государством; за счет современных методов психологической поддержки мотивации. Также должны активно использоваться современные психологические технологии формирования здоровья – методы снятия стресса, методы новой позитивной психологии, методы формирования целеполагания и др.

Принцип надежности, безопасности и конфиденциальности – должен выполняться как для персональных данных пользователя, так и для работы системы в целом, что обеспечивается специальными функциями системы, а также применением при проектировании и реализации системы методик управления рисками. Большое внимание должно уделяться надежности исходных данных, для обеспечения которой применяются специальные методы, в частности, рекомендованные ВОЗ процессы фильтрации и визуализация данных.

Принцип «Эконом – Бизнес - ВИП» Для дальнейшего развития проекта и выхода его на высокую коммерческую эффективность необходимо предусмотреть помимо широкого набора бесплатных сервисов для всего населения, дающих возможность реального улучшения здоровья (эконом-сервисов), набор недорогих платных сервисов для уровня среднего класса (бизнес-сервисы) и, возможно, достаточно дорогих сервисов для ВИП-персон (ВИП-сервисы).

3 Пространство здоровьесбережения

Принцип полноты, требующий представления в проектируемой системе *всех значимых элементов здоровьесбережения*, играет ключевую роль в обеспечении ее эффективности, поэтому необходимо рассмотреть более детально упомянутое выше N-мерное «Пространство Здоровьесбережения», представление которого определяет конкретное содержание и наполнение системы (рис. 1). В соответствии с *принципом открытого эволюционного развития,* наполнение каждого из измерений этого пространства может расширяться до бесконечности, поэтому, в соответствии с *принципом иерархической эффективности*, попытаемся представить наиболее значимые элементы этих измерений, которые целесообразно реализовать в первых версиях системы.

3.1 Показатели здоровья

Целевые критерии: критерии качества жизни, обеспечиваемые здоровьесбережением (характеристики физической и психической работоспособности и эмоционального комфорта); оценки рисков смерти и ожидаемой продолжительности жизни.

Донозологические характеристики здоровья: самооценка здоровья по стандартизированным анкетам, данные обследований.

Нозологические характеристики здоровья: данные МИСов, медицинские карты клиентов, жалобы и персональная информация клиентов.

Показатели, необходимые для обеспечения модулей оптимизации здоровьесбережения: данные функциональных и психологических тестов, специальных опросников, антропометрии.

3.2 Факторы, определяющие здоровье

Факторы: окружающей природной и социальной среды; образа жизни; производственные; риска; генетические и эпигенетические.

3.3 Источники данных о здоровье и определяющих его факторах

МИСы; персональные медицинские карты; биобанки; источники гос. статистики (Роскомстат, Социально-гигиенический мониторинг и др.); м-медицина; интернет вещей; социальные сети и форумы; данные тестирования, самотестирования, анкетирования.

3.4 Методы анализа данных здоровьесбережения

Методы: статистического анализа; визуализации; извлечения данных из соцсетей, форумов, научных текстов; распознавания; классификации; искусственного интеллекта.

3.5 Методы управления здоровьем

Метод управления целями: предоставление клиенту структурированных надежных знаний доступного уровня; методы обеспечения мотивации; методы персонализированной оптимизация программ здоровьесбережения; методы психологической и информационной поддержки реализации программ здоровьесбережения; методы мониторинга эффективности программ и обеспечения обратной связи с клиентом.

Методы интернет-технологий и телемедицины.

3.6 Проблемы здоровья и болезней

Снижающие качество и эффективность жизни: депрессии, стрессы, повышенная утомляемость, раздражительность, психологический дискомфорт, сниженная физическая и психическая работоспособность, сниженная потенция, частые головные и др. боли, ограничения подвижности и боли в скелетно-мышечной системе, проблемы пищеварения.

Медицинские: частые простудные заболевания; хронические заболевания (основные причины смерти) – сердечно-сосудистые, онкологические, эндокринологические, дыхательной системы и др.

3.7 Витальный цикл человека

Зависящие от возраста проблемы здоровья и способы их решения на протяжении всей жизни человека от момента зачатия до конца жизни: здоровое зачатие и вынашивание; формирование здоровья в неонатальном периоде, периодах детства и юношества; проблемы здоровья трудоспособного населения; проблемы здоровья пожилых.

3.8 Все население России (и в перспективе – всего мира)

Различные группы населения, отличающиеся по образованию, доходу, профессии, возрасту, доступу к медицинской помощи, проблемам со здоровьем и т.д.

3.9 Сервисы Системы Здоровьесбережения

Персональный кабинет. Виртуальный и реальный (через телемедицину) персональный врач и тренер здоровья. Надежная структурированная и адаптированная информация о здоровье. Сервисы дистанционного тестирования психо-функционального состояния человека. Сервисы для персонализированной оптимизации оздоровительных программ.

Сервисы обеспечения мотивации и психологической поддержки реализации оздоровительных программ. Сервисы для мониторинга правильности применения и эффективности оздоровительных программ. Мобильное приложение для связи с СЗ. Сервисы анализа и представления обезличенной информации для заказчиков на коммерческих условиях.

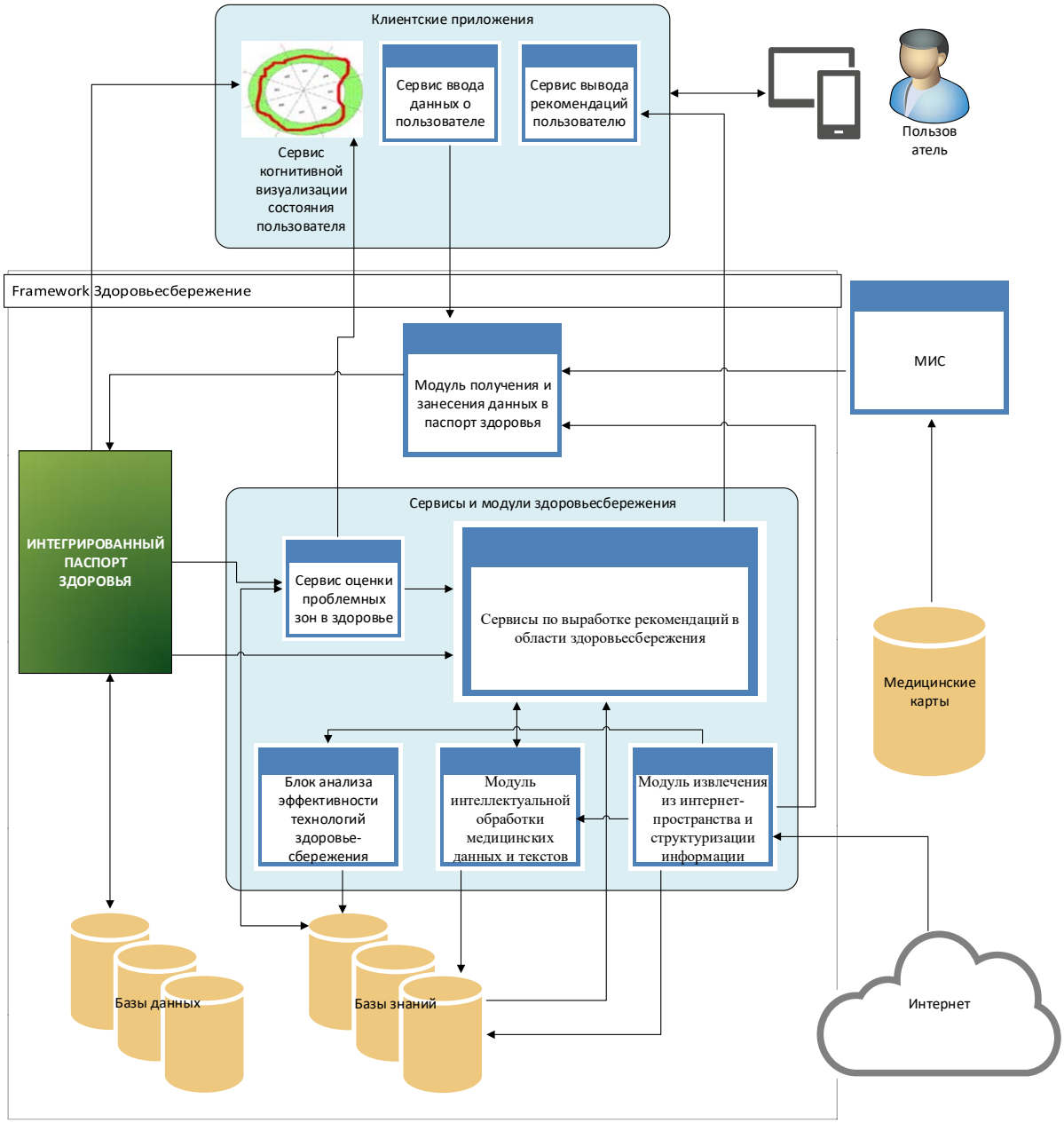


Рисунок 2 Архитектура Системы Здоровьесбережения

4 Архитектура интернет-системы персонализированной поддержки здоровьесбережения

С учетом вышеописанных концептуальных основ здоровьесбережения и структуры его пространства была разработана приведенная архитектура Системы Здоровьесбережения (рисунок 2).

Все основные модули и сервисы расположены на серверной части системы поддержки здоровьесбережения и представляют собой основу (framework), которую можно расширять новыми модулями и сервисами и модифицировать уже имеющиеся.

Модули и сервисы framework объединены в функциональные блоки: блок клиентских приложений; блок анализа и выработки рекомендаций; паспорт здоровья; блок получения и занесения данных в паспорт здоровья; базы данных и знаний.

Клиентские приложения представляют собой ряд сервисов, разворачиваемых на внешних приложениях, с которыми взаимодействует пользователь системы Здоровьесбережения: сервис когнитивной визуализации состояния пользователя, сервисы ввода данных о пользователе, сервисы вывода рекомендаций пользователю. Это базовые сервисы. Каждое клиентское приложение может расширять свой функционал.

Сервис когнитивной визуализации позволяет визуализировать состояние пользователя и графически выделять проблемные зоны в целом и в зависимости от задач, которые хочет решить пользователь с помощью здоровьесбережения. Этот сервис позволяет пользователю оценить свое состояние и его динамику в процессе реализации

программ здоровьесбережения. Также этот сервис используется при выводе данных мониторинга состояния пользователя, визуализации фрагментов рекомендаций, представления характеристик окружающей среды и образа жизни

Сервисы ввода данных о пользователе пересылают информацию о пользователе модулю получения и занесения данных в Интегрированный паспорт здоровья. В качестве способов ввода информации о пользователе выступают: заполнение полей Веб-формы или мобильного приложения; ввод данных с помощью приложений удалённого мобильного мониторинга (mHealth) или систем интернет-вещей (IoT) и многих других способов сбора и передачи данных. Для ввода данных о пользователе системы используется формат Json. Фрагмент Json формата передачи данных о пользователе приведен на рисунке 3.

Сервис вывода рекомендаций пользователю. В зависимости от проблем со здоровьем, решением которых планирует заняться пользователь системы, ему предлагаются различные персонализированные рекомендации, сгенерированные модулями оптимизации программ здоровьесбережения. Данные рекомендации могут иметь различный вид – от рекомендаций ознакомиться с хранящейся в системе научной информацией по проблемам пользователя до детального описания персонализированной программы здоровьесбережения.

```
{
   "patient_psv": {
      "ch_number": "123",
      "value": "120",
      "date_time": "15.02.2015 18:25"
   }
}

{
 "patient_state":{
"zone":"зеленая"
}
```

Рисунок 3 Фрагмент формата передачи данных

Модуль получения и занесения данных в Интегрированный паспорт здоровья. Основное назначение этого модуль – обеспечение сбора информации о пользователе системы из различных источников и запись ее в паспорт здоровья. В качеств источников информации выступают: сам клиент (ручной или автоматизированный ввод данных из тест-опросников, систем дистанционного тестирования, приборов mHealth); истории болезней, хранящиеся в базах данных медицинских информационных систем (МИС) лечебных учреждений; социальные сети и форумы; базы данных Росстата, Социомониторинга и др. государственных информационных систем, которые имеют общую информацию о показателях окружающей социальной и природной среды. Данный модуль должен позволять передавать информацию, полученную с помощью практически любых способов ее сбора. Это свойство необходимо для снятия всех ограничений на подключение к Системе Здоровьесбережения новых сервисов и программ, которые будут появляться извне или специально разрабатываться при появлении новых проблем здоровьесбережения.

Сервисы и модули здоровьесбережения.

В данном блоке присутствуют следующие основные модули и сервисы: сервис оценки проблемных зон в здоровье; подблок модулей для выработки рекомендаций по персонализированному здоровьесбережению; модуль интеллектуальной обработки медицинских данных и текстов; модуль извлечения из интернет-пространства и структуризации информации; блок анализа эффективности технологий здоровьесбережения.

Модуль оценки проблемных зон в здоровье объективизирует проблемы здоровья пользователя на основе анализа данных Интегрированного паспорта здоровья (определяет риски заболеваний и смерти, уровни физической и психической работоспособности и т.д.), передает оценку этих проблем как в модуль Когнитивной визуализации для представления пользователю его же собственных проблем в наглядном виде, стимулирующем их решение, так и в Подблок выработки рекомендаций по персонализированному здоровьесбережению.

Подблок выработки рекомендаций по персонализированному здоровьесбережению включает в себя модули, генерирующие оптимальные персональные рекомендации по основным направлениям ЗОЖ, превращая его таким образом в ЗОЖ-HiTech. В совокупность этих модулей входят:

Модуль оптимизации рациона питания, учитывающий привычки и предпочтения пользователя и устраняющий дефициты эссенциальных нутриентов в рационе (данный модуль, как и многие другие в системе будет давать рекомендации разной детальности и стоимости для пользователя (от Эконом до ВИП уровней), где ВИП уровень будет использовать, например, ОМИКсные данные пользователя).

Модуль оптимизации физической активности, рекомендующий различные виды физической активности, в зависимости от предпочтений пользователя, и уровни нагрузок, в зависимости от его физических кондиций и типов нагрузок.

Модуль для выработки программ улучшения психической работоспособности, помогающий оптимизировать программы тренировки психофизиологических функций и циркадный график интеллектуальной работы.

Модуль для выработки рекомендаций по коррекции психоэмоционального состояния: снятие стресса, применение медитативных техник, улучшение мотивации на здоровьесбережение, психологическая поддержка процессов здоровьесберегающей коррекции образа жизни.

Модуль для выработки рекомендаций по здоровьесберегающему образу жизни, основанных на результатах анализа качества жизни.

Модуль для коррекции рисков профессиональной деятельности, дающий оценку степени опасности для здоровья имеющихся рисков и структурированную информацию, помогающую их корректировать.

Модуль для организации дистанционных персональных on-line консультаций пользователя с врачами и тренерами.

Модуль поддержки осуществления лечебных и восстановительных процедур, рекомендованных врачами и тренерами в связи с наличием у пациента определенных заболеваний или их рисков.

Отдельную важную роль играет входящий в Блок анализа и выработки рекомендаций *Модуль анализа эффективности технологий здоровьесбережения*. Данный модуль предназначен для получения новых данных и знаний об эффектах действия и безопасности различных средств и методов коррекции здоровья на основе анализа больших данных из пространства интернет – из социальных сетей, форумов и научных текстов. Данный модуль является основой дальнейшего развития и совершенствования СЗ.

Модуль извлечения из интернет-пространства и структуризации информации предназначен для обхода сети Интернет и извлечения информации как о пользователях различными методами здоровьесбережения, так и об эффективности этих методов. В качестве источников информации выступают социальные сети, тематические форумы и блоги, базы данных и сервисы научных публикаций. Этим модулем извлекается как информация о пользователе системы, его отношении к различным технологиям, эффективности применения тех или иных здоровьесберегающих технологий, в каких форумах и группах он состоит и др., так и общая информация о здоровьесберегающих технологиях, которая извлекается из различных публикаций. Вся извлекаемая информация передается в блок анализа эффективности технологий здоровьесбережения и в модуль интеллектуальной обработки медицинских данных и текстов.

Модуль интеллектуальной обработки медицинских данных и текстов.

Данный модуль служит для извлечения знаний об эффективности различных средств и методов здоровьесбережения из медицинских данных и текстов, а также для оценки значимости различных факторов, детерминирующих здоровье.

Базы данных и знаний, присутствующие в системе содержат разнообразную информацию, необходимую для реализации функций всех блоков Системы Здоровьесбережения: данные о характеристиках здоровья пациентов и определяющих их здоровье факторах; официальные нормативы и стандарты для величин данных характеристик и факторов; структурированная информация по здоровьесбережению из научных текстов и медицинских рекомендаций; описания наиболее эффективных средств и методов здоровьесбережения и т.д.

Особую роль играет база данных, содержащая информацию о характеристиках здоровья конкретных пациентов и определяющих его здоровье факторах – база данных ***Интегрированных паспортов здоровья***. Термин «интегрированный» подчеркивает важность наличия в паспорте здоровья не только характеристик здоровья, но и определяющих здоровье факторов, что отличает этот паспорт от традиционных паспортов здоровья, используемых в медико-профилактической сфере.

Заключение

В работе рассмотрена медико-демографическая проблема России, представлены концептуальные основы (основные принципы построения и требования) системы здоровьесбережения, описана структура пространства управления здоровьем и общая архитектура интернет-системы персонализированной поддержки здоровьесбережения. Основными функциями системы являются: сбор информации из различных источников, включая интернет-пространство; интеллектуальная обработка медицинских данных и текстов; анализ эффективности технологий здоровьесбережения; оценка проблемных зон в здоровье конкретного человека; персонализированная оптимизация технологий его здоровьесбережения; помощь в их применении и мониторинг их эффективности.

Ядро этой системы содержит Интегрированный паспорт здоровья и расширяемый набор модулей и сервисов, позволяющих собирать информацию из различных источников, обрабатывать и выдавать пользователю персонализированные рекомендации о применении профилактических мер по улучшению/стабилизации его состояния здоровья. Особой частью ядра является Интегрированный паспорт здоровья, структура которого может меняться или дополняться без влияния на другие составные части системы поддержки здоровьесбережения. Поэтому же принципу спроектированы и остальные модули, блоки и сервисы. Вся архитектура разработана таким образом, чтобы к ней можно было подключать сторонние приложения, к которым планируется добавлять сервисы по здоровьесбережению. Интегрированный паспорт здоровья может быть использован и как база биобанка – возможность выбора репрезентативных групп обследуемых для участия в испытаниях новых эффективных средств здоровьесбережения.

Литература

[1] Глобальная стратегия по питанию, физической активности и здоровью. Всемирная организация здравоохранения. 2004. [Электронный ресурс]// http://www.who.int/publications/list/9241592222/ru/

[2] Государственная программа Российской Федерации "Развитие здравоохранения". 2014. [Электронный ресурс]// https://www.rosminzdrav.ru/ministry/programms/health/info

[3] Донцов В.И., Мамиконова О.А., Потемкина Н.С., Смирнова Т.М. Концепция и архитектура интегрального паспорта здоровья // Вестник восстановительной медицины. 2016. № 1. С.14-20.

[4] Донцов В.И., Крутько В.Н. Здоровьесбережение как современное направление профилактической медицины (обзор) // Вестник восстановительной медицины. 2016. № 1. С.2-9.

[5] Прогноз долгосрочного социально-экономического развития Российской Федерации на период до 2030 года. Минэкономразвития России, Москва, март 2013. [Электронный ресурс] // government.ru/media/files/41d457592e04b76338b7.pdf.

[6] Резолюция круглого стола «Изменяя мир: укрепление здорового образа жизни и контроля за неинфекционными заболеваниями» Генеральной ассамблеи ООН, Нью-Йорк, 2015 г. [Электронный ресурс]// https://www.rosminzdrav.ru/news/2015/09/25/2550-ministr-veronika-skvortsova-vystupila-na-kruglom-stole-izmenyaya-mir-ukreplenie-zdorovogo-obraza-zhizni-i-kontrolya-za-neinfektsionnymi-zabolevaniyami

[7] David E. Bloom, David Canning, Jaypee Sevilla. The Effect of Health on Economic Growth: Theory and Evidence. NBER Working Paper No. 8587. November 2001. [Электронный ресурс] // http://www.nber.org/papers/w8587

[8] Global Health Observatory Data Repository. Life expectancy. [Электронный ресурс] // http://apps.who.int/gho/data/node.main.687?lang=en

Conceptual foundation and architecture of the Internet system for personalized healthcare support using data intensive analysis

Vyacheslav N. Krutko, Alexey I. Molodchenkov

The problem of health of Russian population and its solutions are described in the article, objectives are formulated and conceptual foundation of healthcare system, structure of health management and architecture of an Internet system for the personalized healthcare support are described. The main functions of a system are: to collect information from various sources, including the Internet space; intelligent processing of medical data and text; to analyze healthcare technologies effectiveness; to estimate problem areas in the health of a particular person; personalized optimization of healthcare technologies, assistance of its application and monitoring of its effectiveness.

The reported study was funded by Ministry of Education and Science of Russia according to the research project № 14.607.21.0123.

Использование компьютерной системы оценки психической работоспособности в режиме домашней лаборатории

© Т. М. Смирнова © В. Н. Крутько

Институт системного анализа ФИЦ ИУ РАН,
Москва
smirnova.tatyana@gmail.com krutkovn@mail.ru

Аннотация

Компьютерная система «СОПР-мониторинг» была разработана нами для целей психофизиологического мониторинга в экспериментальных исследованиях в области космической медицины. Этот компьютерный инструмент был использован в разнообразных экспериментах, в том числе в международном проекте «Марс-500». В домашних условиях система «СОПР-мониторинг» применялась в целях многолетнего индивидуального мониторинга умственной работоспособности, сенсомоторной координации и психо-эмоционального состояния. Анализ данных этого исследования позволил выявить ряд индивидуальных особенностей в динамике психо-эмоционального состояния и работоспособности: долговременные тренды, суточные и годичные ритмы, чувствительность к экстремальным погодным условиям. Результаты индивидуального мониторинга могут быть использованы в целях оптимизации режима труда и отдыха, в особенности в условиях удаленной работы.
Исследование выполнено при финансовой поддержке Минобрнауки РФ в рамках проекта № 14.607.21.0123.

1 Цели и методы исследования

Исследование было проведено в целях проверки возможностей использовать компьютерную систему «СОПР-мониторинг», эффективность которой была многократно подтверждена в экспериментах по изучению состояния и работоспособности человека в условиях реальных и моделированных факторов космического полета, для оценки индивидуальных характеристик состояния и работоспособности при выполнении привычной профессиональной деятельности в домашних условиях.

Труды XVIII Международной конференции DAMDID/RCDL'2016 «Аналитика и управление данными в областях с интенсивным использованием данных», Ершово, 11-14 октября 2016

Компьютерная система «СОПР-мониторинг» [1] была разработана при участии сотрудников ГНЦ РФ - ИМБП РАН – ведущей российской организации в сфере медико-биологического обеспечения пилотируемых космических объектов и фундаментальных исследований в области космической биологии и медицины. «СОПР-мониторинг» представляет собой компактную компьютерную реализацию ряда тестовых методик, которые традиционно используются в космической медицине и показали высокую информативность. Ранее некоторые из этих методик имели аппаратную реализацию, а анализ психологического состояния космонавтов и испытателей-добровольцев осуществляли с помощью анкетирования в бумажном виде.

Управление работой системы «СОПР-мониторинг» организовано по принципу меню. На каждом этапе работы на монитор выводятся кнопки управления для выбора возможных вариантов действий пользователя и подсказки, описывающие возможные действия.

Перед началом первого сеанса работы пользователь должен зарегистрироваться, указав фамилию, имя, отчество, дату рождения и пол. Два поля – «Профессии» и «Виды спорта», - отражающие имеющийся у пользователя опыт, не обязательны для заполнения, однако содержащаяся в них информация может оказаться важной для интерпретации результатов мониторинга.

После первичной регистрации или выбора нужного имени в списке зарегистрированных пользователей последовательно предъявляется следующий набор тестовых методик:

- 8-цветовой тест Люшера;
- тест САН (самооценка самочувствия, активности и настроения);
- непрерывный счет в автотемпе (НСАТ);
- реакция на движущийся объект (РДО);
- тест Спилбергера-Ханина на реактивную тревожность;
- тест Люшера повторно.

Таким образом, набор тестовых методик обеспечивает оценку наиболее лабильных

компонент психической работоспособности – умственной работоспособности и сенсомоторной координации, а также оценку психо-эмоционального состояния. Реализация теста САН в системе «СОПР-мониторинг» несколько отличается от традиционного бланкового метода. Как и в бланковом варианте, последовательно предъявляется 30 пар контрастных утверждений относительно состояния пользователя, но шкала оценки близости состояния к одному из крайних вариантов – не дискретная с 3 градациями с каждой стороны, а непрерывный интервал от -1 (наихудшее возможное состояние) до 1 (наилучшее состояние). По 30 ответам вычисляются интегральные оценки по всем трем шкалам теста, которые заносятся в файл результатов.

В процессе выполнения методики НСАТ в течение 3 минут на монитор последовательно выводятся однозначные числа зеленого, красного или синего цвета. Пользователь должен, начиная с исходного числа, в дальнейшем все зеленые числа прибавлять, а красные – вычитать из предыдущей суммы, набирая результат на цифровой клавиатуре. В случае появления синего числа необходимо повторить предыдущий результат. По результатам этого теста фиксируются общее число выполненных операций и число правильных ответов, а также рассчитывается показатель точности счета – процентная доля правильных ответов.

После вызова теста РДО на экране появляется окружность с неподвижной меткой и шариком, который начинает двигаться по окружности. Задача пользователя – нажимать на клавишу <Enter> каждый раз, когда подвижная метка касается неподвижной. Продолжительность теста – около 3 минут, за это время шарик проходит 30 кругов. В файле результатов фиксируется время ошибки реакции для каждого круга (если на каком-то круге клавиша <Enter> не была нажата, - пробел), а также расчетные величины – средний модуль ошибки (СМО), средняя величина ошибки (СО) и максимальный модуль ошибки (ММО).

В тесте Спилбергера-Ханина пользователь должен выбрать один из четырех предлагаемых вариантов ответа на 20 последовательно задаваемых вопросов. Результатом теста является интегральная оценка реактивной тревожности.

Система «СОПР-мониторинг» применялась в разнообразных экспериментах, проведенных на базе ГНЦ РФ - ИМБП РАН [2, 4, 5, 7] и посвященных исследованию индивидуальной и среднегрупповой динамики психо-эмоционального состояния и работоспособности в малых группах испытателей–добровольцев на фоне имитации комплекса факторов космического полета. Продолжительность экспериментов соответствовала длительности реальных и проектируемых космических миссий – от нескольких суток до 520 суточного эксперимента по моделированию пилотируемого полета на Марс с высадкой на поверхность планеты, в котором участвовали 6 испытателей из России, Италии, Франции и Китая.

В целях исследования возрастной динамики и половых различий психо-эмоционального состояния и умственной работоспособности система «СОПР-мониторинг» была использована при обследовании профессионально однородной группы из 10 мужчин и 17 женщин в условиях привычной профессиональной деятельности [6].

Также эта система была использована в клиническом исследовании в рамках межведомственного проекта «Живи долго!» [3]. Целью данного проекта, выполненного на базе ГКБ № 11 города Воронежа, являлись исследование динамики старения и апробация методов замедления старения в условиях современного российского города. В рамках этого проекта были обследованы лица пожилого возраста (51 - 87 лет; всего 80 человек), поэтому наиболее сложная нагрузочная проба – тест НСАТ – не использовалась. С прочими тестами обследуемые успешно справились, что доказывает удобство системы тестирования и ее приемлемость для лиц со значительно более низкими функциональными возможностями, чем у космонавтов и испытателей, допущенных к участию в экспериментах в области космической медицины.

В домашних условиях система «СОПР-мониторинг» была использована для многолетнего (с марта 2007 г. по май 2016 г., всего 190 сеансов) тестирования одним человеком. Это исследование не имело заранее заданного плана, поэтому интервалы между сеансами варьировали от неполных суток до 286 дней. Тестирование осуществлялось в разное время суток в перерывах в процессе работы за компьютером – либо выполнения привычной профессиональной деятельности, либо поиска информации для различных индивидуальных целей. В ходе домашнего тестирования была использована возможность заносить в файлы результатов комментарии, уточняющие условия тестирования.

Для анализа результатов домашнего мониторинга были использованы методы корреляционного, регрессионного и дисперсионного анализа. Вычисления выполнены с помощью пакета Statistica.

2 Результаты и обсуждение

2.1 Долговременные тренды

Анализ многолетних тенденций изменения состояния организма и работоспособности представляет интерес в первую очередь в плане выявления возрастных трендов. В таблице 1 приведены значения коэффициентов корреляции между выборками показателей мониторинга и рядом моментов наблюдений (за начало отсчета было принято начало первого сеанса). Достоверное снижение качества деятельности было отмечено по всем показателям теста НСАТ, но не по показателям теста РДО. Самооценка состояния ухудшалась со

временем по всем параметрам, однако значимых изменений тревожности не наблюдалось. Вклад линейной временной компоненты в динамику показателей мониторинга был невелик, - коэффициент детерминации ни в одном случае не достигал 20%. Линейная аппроксимация даже наиболее тесно коррелированных со временем показателей по формуле:

$$y=A+B*t,$$

где y – показатель мониторинга,
t - время,
A и B – коэффициенты,

показала, что среднегодовое снижение показателей умственной работоспособности и самооценки состояния не превышало нескольких процентов от начального уровня (таблица 2).

Таблица 1 Корреляционные связи показателей мониторинга со временем регистрации.

Показатели	r	r^2, %	p
Число правильных ответов	-0,427	18,3	0,000
Число операций	-0,423	17,9	0,000
Точность	-0,205	4,2	0,005
СМО	-0,050	0,2	0,498
СО	0,043	0,2	0,559
ММО	-0,071	0,5	0,330
Самочувствие	-0,254	6,4	0,001
Активность	-0,178	3,2	0,016
Настроение	-0,374	14,0	0,000
Тревожность	-0,033	0,1	0,648

Таблица 2 Параметры линейного тренда показателей мониторинга.

Показатели	A	B, 1/год	B*1год, % A
Число правильных ответов	119,5	-1,3	-1,1
Число операций	119,9	-1,3	-1,0
Точность, %	99,6	-0,1	-0,1
Самочувствие	0,59	-0,02	-2,6
Активность	0,34	-0,01	-2,6
Настроение	0,51	-0,02	-3,6

Однако при более детальном рассмотрении динамики показателей мониторинга можно увидеть, что связь с фактором времени была значимой и для показателей сенсомоторной реакции (таблица 3). Это свидетельствует о немонотонном характере динамики этих показателей. Не были монотонными изменения даже наиболее тесно коррелированных со временем показателей умственной работоспособности (рисунки 1 и 2). В отдельные годы наблюдался рост этих показателей, что свидетельствует о наличии иных, помимо возрастных изменений, факторов, влияющих на уровень умственной работоспособности.

Таблица 3 Уровни значимости связи показателей мониторинга с годом наблюдения.

Показатели	p
Число правильных ответов	5,2E-08
Число операций	6,9E-08
Точность	3,9E-03
СМО	1,1E-07
СО	2,4E-04
ММО	7,1E-06
Самочувствие	4,5E-04
Активность	7,5E-05
Настроение	5,9E-07
Тревожность	0,371

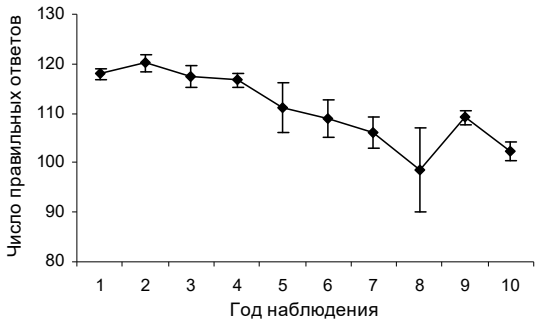


Рисунок 1 Динамика числа правильных ответов по годам наблюдения

Примечание: на этом и всех следующих рисунках приведены средние и стандартные ошибки среднего.

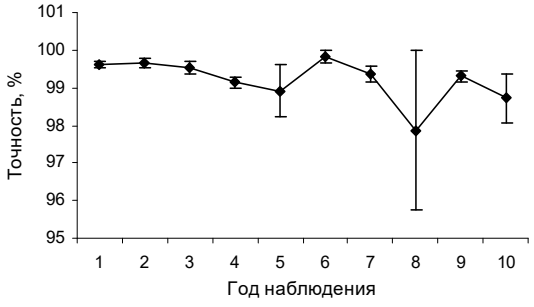


Рисунок 2 Динамика точности счета по годам наблюдения.

Показатели ошибки сенсомоторной реакции снижались не только в первые 2-3 года тестирования (что можно было бы связать с эффектом обучения), но и в дальнейшем – на шестой год по сравнению с

пятым (рисунки 3 и 4). Монотонный рост СМО и СО (но не ММО), который можно было бы связать с возрастными изменениями, имел место, начиная только с седьмого года наблюдения.

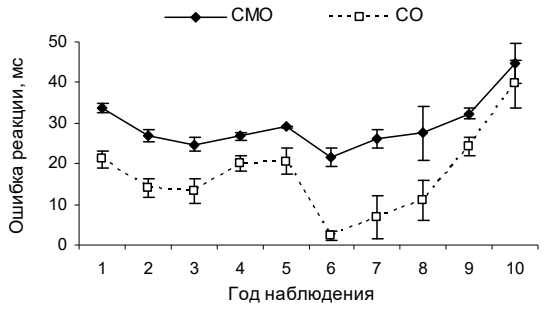


Рисунок 3 Динамика среднего модуля и средней ошибки реакции по годам наблюдения.

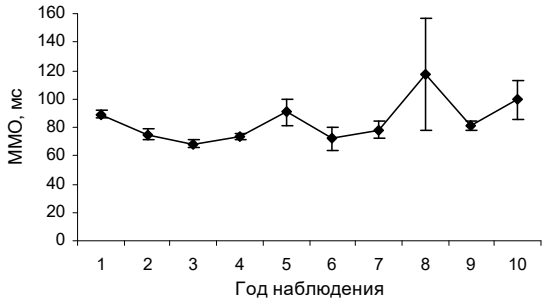


Рисунок 4 Динамика максимального модуля ошибки реакции по годам наблюдения.

Динамика показателей самооценки состояния также не была монотонной, причем вид всех трех кривых был различен (рисунки 5-7).

Различному характеру временной динамики показателей мониторинга не может быть дано однозначное объяснение на основе имеющейся информации об условиях исследования на всем его протяжении. Отсюда следует, что при проведении мониторинга вне строго контролируемых условий научного эксперимента необходимо более подробное описание условий тестирования.

2.2 Суточная периодичность

Анализ суточных периодичностей работоспособности важен для определения оптимального режима труда и отдыха.

Для анализа суточных периодичностей время выполнения тестов было категоризировано по двухчасовым интервалам. Результаты дисперсионного анализа связей между показателями мониторинга приведены в таблице 4. Почти для всех показателей, за исключением точности счета и настроения, эта связь оказалась статистически значимой.

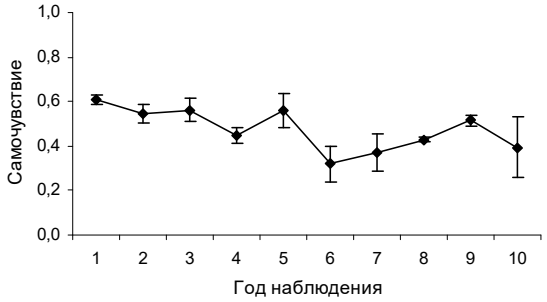


Рисунок 5 Динамика самочувствия по годам наблюдения.

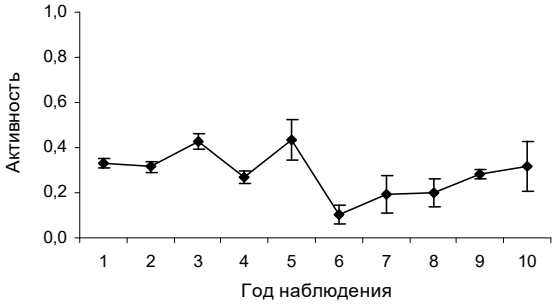


Рисунок 6 Динамика активности по годам наблюдения.

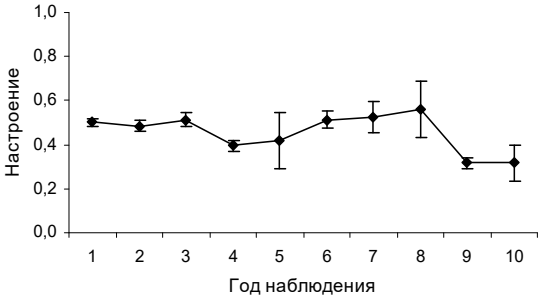


Рисунок 7 Динамика настроения по годам наблюдения.

Таблица 4 Уровни значимости связи показателей мониторинга со временем суток.

Показатели	p
Число правильных ответов	0,050
Число операций	0,043
Точность	0,094
СМО	0,004
СО	0,002
ММО	0,001
Самочувствие	0,015
Активность	0,000
Настроение	0,275
Тревожность	0,008

Умственная работоспособность была значительно ниже в ночное время по сравнению с дневными и вечерними часами Максимальная работоспособность имела место в период от 8 до 10 часов утра, после чего довольно плавно снижалась вплоть до полуночи (рисунок 8).

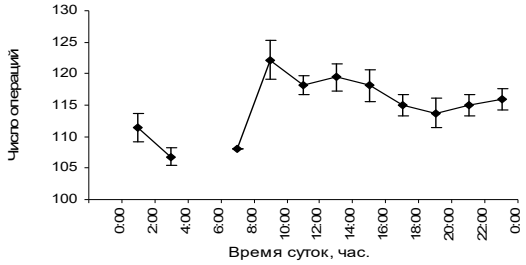


Рисунок 8 Суточная периодичность числа операций.

Максимальная эффективность сенсомоторной деятельности также отмечалась между 8 и 10 часами утра, но затем ошибка реакции резко возрастала (рисунок 9). Во второй половине дня сенсомоторная координация улучшалась, и между 22 и 24 часами имел место второй минимум ошибки. Таким образом, точки наилучшей и наихудшей работоспособности для двух видов деятельности – умственной и сенсомоторной – неодинаковы.

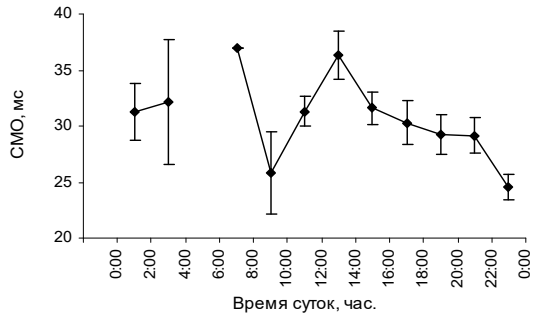


Рисунок 9 Суточная периодичность среднего модуля ошибки.

Суточная динамика самочувствия не имела сколько-нибудь выраженного закономерного характера (рисунок 10). Максимум в интервале 6-8 часов был получен за счет единственного сеанса, пришедшегося на этот интервал времени, и не может считаться закономерным.

Среди показателей самооценки состояния наиболее выраженная циркадианная динамика имела место для показателя активности (рисунок 11). Если исключить из рассмотрения единственное наблюдение в интервале 6-8 часов, то максимум активности приходится на 8-10 часов, т.е. совпадает с максимумами эффективности как умственной, так и сенсомоторной деятельности. Период выраженного спада активности соответствует снижению качества сенсомоторной координации. В то же время максимальные уровни тревожности соответствуют периоду максимальной умственной работоспособности (рисунок 12), что, вероятно, является проявлением напряжения, связанного с этим видом деятельности. Это предположение согласуется с тем фактом, что тестирование выполнялось на фоне работы за компьютером, характер которой в значительно большей степени связан с интеллектуальной нагрузкой, чем с сенсомоторной. Результаты исследования суточных ритмов позволяют оптимальным образом организовать распределение разнохарактерных видов деятельности в течение суток.

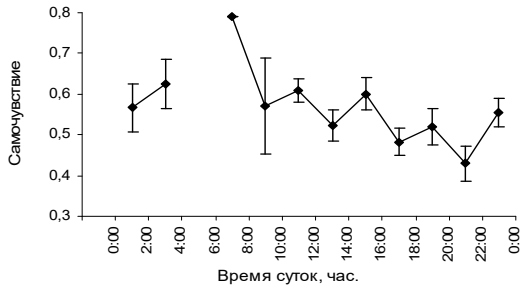


Рисунок 10 Суточная периодичность самочувствия.

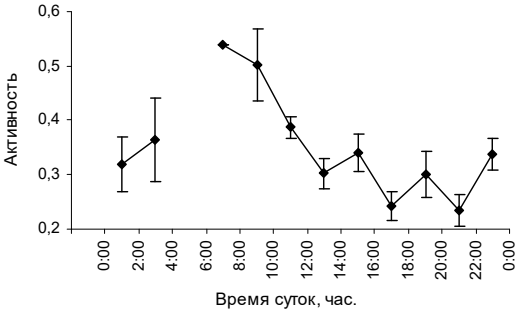


Рисунок 11. Суточная периодичность активности.

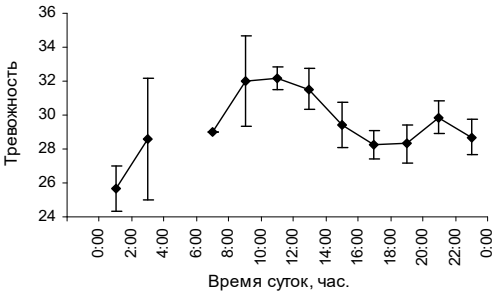


Рисунок 12 Суточная периодичность тревожности.

2.3 Сезонная периодичность

Наличие цирканнулярных (т.е. окологодичных) ритмов в динамике многих физиологических и психологических показателей может сказываться и на показателях работоспособности. В связи с этим была исследована динамика показателей мониторинга по месяцам года. Как видно из таблицы 5, связь с фактором «месяц года» была значимой для всех показателей эффективности деятельности, кроме точности счета, но не значимой – для показателей психологического состояния.

Колебательный характер помесячной динамики умственной работоспособности не позволяет выявить какой-либо явной закономерности (рисунок 13). В динамике СМО и СО выражено резкое повышение в весенние месяцы (рисунок 14). Самый низкий уровень эти показатели имели в период с

сентября по декабрь. Такой результат позволяет предположить, что для данного обследуемого целесообразно было бы по возможности переносить на этот период года работы, требующие высокой точности реакции.

Таблица 5 Уровни значимости связи показателей мониторинга с месяцем наблюдения.

Показатели	p
Число правильных ответов	0,025
Число операций	0,036
Точность	0,400
СМО	5,24E-10
СО	9,05E-10
ММО	2,07E-05
Самочувствие	0,084
Активность	0,644
Настроение	0,137
Тревожность	0,437

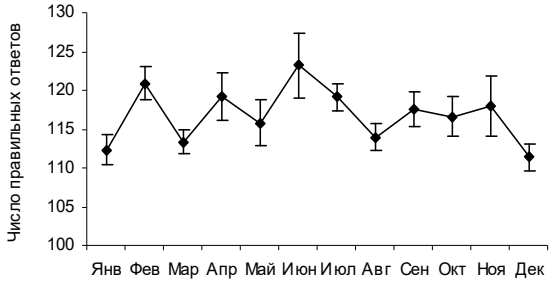


Рисунок 13 Среднемесячные уровни числа правильных ответов.

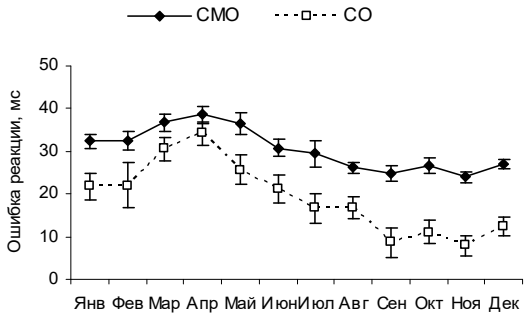


Рисунок 14 Среднемесячные уровни ошибки реакции.

Возможно, что слабо выраженная годичная периодичность умственной работоспособности связана с тем, что для данного испытателя этот вид деятельности является основным. Многолетняя тренировка, вероятно, подавляет вариабельность показателей привычного вида деятельности.

2.4 Чувствительность к экстремальным погодным условиям

В основном настоящее исследование проходило в комфортных погодных условиях. Однако на четвертый год мониторинга пришелся период экстремальной жары и загрязненности воздуха летом 2010 г. Для оценки чувствительности показателей мониторинга к этому природному воздействию мы сравнили средние значения показателей за весь четвертый год наблюдения (чтобы исключить влияние сезонной вариабельности) в период жары и вне этого периода. Чувствительными к данному воздействию оказались только два показателя – точность счета и настроение (таблица 6). При этом точность в период жары была ниже, но настроение – выше, чем в обычных погодных условиях. Повышение настроения, не сопровождающееся улучшением самочувствия, отражает, по-видимому, не столько индивидуальную переносимость жары, сколько индивидуальный механизм мобилизации в неблагоприятных условиях, обеспечивающей устойчивый уровень почти всех мониторируемых показателей работоспособности.

Таблица 6 Показатели мониторинга, чувствительные к экстремальной жаре.

Показатели	Период жары	4-й год без периода жары	p
Точность, %	98,8±0,2	99,5±0,2	0,017
Настроение	0,46±0,03	0,34±0,03	0,013

Заключение

Результаты мониторинга с помощью системы «СОПР-мониторинг» позволили выявить ряд индивидуальных особенностей работоспособности и психо-эмоционального состояния, которые имеет смысл учитывать в целях оптимального планирования режима труда и отдыха. Это особенно значимо для работающих в дистанционном режиме.

Система «СОПР-мониторинг» может быть использована в домашних условиях для тренировки умственной работоспособности и сенсомоторной деятельности.

Пользователь, достаточно подготовленный в области планирования эксперимента и анализа данных, может с помощью данного компьютерного инструмента организовать для себя и членов своей семьи исследование показателей психической работоспособности при любых условиях и воздействиях, представляющих интерес, и разрабатывать оптимальные режимы труда и отдыха самостоятельно.

Блок индивидуального мониторинга психической работоспособности и психо-эмоционального состояния может быть включен в интернет-среду, обеспечивающую, с одной стороны, статистический анализ данных, накопленных отдельными пользователями, и передачу результатов анализа пользователям, а с другой – накопление и анализ больших объемов обезличенных данных. Такой вариант организации мониторинга и оптимизации режимов деятельности на основе его результатов может быть реализован в рамках разрабатываемой в ИСА РАН интернет-системы персонализированной

поддержки здоровьесбережения через личный кабинет пользователя. Данный подход позволил бы использовать результаты индивидуального мониторинга как в целях индивидуального и семейного здоровьесбережения, так и в качестве элемента системы социально-гигиенического мониторинга, обеспечивающего оценку работоспособности и психо-эмоционального состояния на популяционном уровне.

Литература

[1] Т. М. Смирнова, В. Н. Крутько, А. Ф. Быстрицкая, А. Г. Виноходова, И. М. Ларина. Использование компьютерной системы «СОПР-мониторинг» для анализа психической работоспособности в условиях привычной профессиональной деятельности и в сложных условиях. Труды Института системного анализа Российской академии наук, т. 19, 2006, с. 156-170.

[2] А. Г. Виноходова, Т. М. Смирнова, А. Ф. Быстрицкая, В. Н. Крутько. Использование компьютерных методов оценки «СОПР-мониторинг» для оценки работоспособности при моделировании факторов космического полета. Авиакосмическая и экологическая медицина, 41(6), 2007, с. 48-52.

[3] В. Н. Крутько, Т. М. Смирнова, М. В. Силютина, О. Н. Таранина. Психологические и клинико-физиологические корреляты старения у женщин. Вестник восстановительной медицины, № 2, 2015, с. 2-6.

[4] А. Г. Виноходова, А. Ф. Быстрицкая, Т. М. Смирнова. Способность к психической саморегуляции как фактор устойчивости к стрессу в экстремальных условиях космического полета. Авиакосмическая и экологическая медицина, 39(5), 2005, с. 14-18.

[5] В. И. Гущин, Д.М. Швед, М. А. Левинских, А. Г. Виноходова, О. Б. Сигналова, А. Е. Смолеевский. Экопсихологические исследования в условиях 520-суточной изоляции. Авиакосмическая и экологическая медицина, 48(3), 2014, с. 25-29.

[6] V.N.Krutko, T.M.Smirnova, I.M.Larina, G.Y.Vasilieva. Mental performance and sensomotoric coordination as markers of bioage: sexual differences, nonlinear effects, phsychophysiological correlations. VI European Congress: Healthy and Active Ageing for All Europeans. July 5-8, 2007, Saint Petersburg, Russia. Advances in Gerontology, 2007, volume 20, № 3, p. 49.

[7] D. M. Shved, V. I. Gushin, A.G. Vinokhodova, I. A. Nichiporuk, G. Y. Vasilieva. Psychophysiological adaptation and communication behavior of a human operator during 105-day isolation. Human Physiology, 40(7), 2014, p. 797-803.

Utilization of computerized evaluation of mental performance in home lab mode

Tatyana M. Smirnova, Vyacheslav N. Krutko

Computer system "SOPR-monitoring" has been developed for the purposes of psycho-physiological monitoring in experimental research in the field of space medicine. This computer-based tool was used in various experiments, including the international project "Mars-500". At home, "SOPR-monitoring" was applied to multi-year individual monitoring of mental performance, sensomotoric coordination and psycho-emotional state. Analysis of the data from this study revealed a number of individual features in dynamics of psycho-emotional state and performance: long-term trends, daily and annual rhythms, sensitivity to extreme weather conditions. The results of individual monitoring can be used in order to optimize the work/rest schedules, especially in the context of remote job.

The reported study was funded by Ministry of Education and Science of Russia according to the research project № 14.607.21.0123.

Using the DTW method for estimation of deviation of care processes from a care plan

© Alexey Molodchenkov © Mikhail Khachumov

Federal Research Center "Computer Science and Control" of Russian Academy of Sciences, Moscow

aim@isa.ru, khmike@inbox.ru

Abstract

Hospitals increasingly use process models for structuring their care processes. Activities performed to patients are logged to a database or a log. These data can be used for managing and improving the efficiency of care processes and quality of care. In this article, we propose the method for estimation of deviation of care processes from a care plan. Care plan defines the steps of a patient treatment for a certain disease in a specific hospital. Care plan is built on the base of care process model. A care process model is built on the base of exemplars of care processes, stored in a database or a log. The Dynamic Time Warping (DTW) algorithm was used for estimation of deviation. The DTW algorithm measures a distance-like quantity between two given traces containing information about execution or not execution of actions defined by care plan.

1 Introduction

Explore of methods for management of care process (CP) as a flow of therapeutic and diagnostic activities allows us to analyze situations with patients and recommend to decision-makers (DM) appropriate action for the treatment of patients. In this connection, occurs increased interest in an automation of hospitals units (accounting, reception, offices, warehouse and so on) and an analysis and a formal representation of care processes in order to support doctors work.

The first area is deep enough elaborated - powerful medical information systems for supporting of therapeutic and diagnostic processes are developed. Flows of patients can be optimized by simulation in order to identify bottlenecks.

The second area requires additional researches to estimate stages of a care process and to develop recommendations for decision-makers. Currently tools for analysis and medical diagnostics, using different classifiers with high accuracy and precision are developed [3, 4, 8, 11, 17, 18, 24]. This area has a number of features and is of great interest for us. A care process, despite the standards, always has an individual (personalized) character. There may be various deviation from the selected care plan, depending on the changing care conditions, concomitant deceases, etc. Therefore, there are not only problems replying of a care process, and operational management of a care process in terms of possible deviations. Management is a particular sequence of treatment actions (operations), which is based on states of a patient, a prescribed care plan and medical databases, i.e., precedents.

The aim of our research is to develop algorithms and tools that assist to a doctor by making proposals (recommendations) on the organization of care process in accordance with an actual patient's state and care plan. One of the objectives is to assess the quality of the rendered medical services by comparing a patient's care process and care plan. The care plan defines the actions of the doctor and is based on the care process model, discovered from precedents.

The multi-dimensional distance based on Dynamic Time Warping (DTW) algorithm is used to differences between care process and care plan. Experiments show that DTW algorithm is effective to compare sequences of actions in relation to care processes. We consider the use of DTW algorithm to detect deviations between real care process of a patient and care plan.

2 Formalization and optimization of care processes

Some methods for formalization and application of care processes are described in [1, 12, 17]. Methods and algorithms for discovery of models of care processes on the base of event logs and precedents are described in [3, 6, 12-16, 19]. The model includes all traces from event logs and precedents. Fig. 1 illustrates the Petri net workflow process definition for handling a medical complaint. In this figure we can identify the following routing constructs: transitions "Identification" and "Cardiologist" are AND-splits, "I diagnosis OK", "I diagnosis NOK" and "decide surgery" are AND-joins, c4, c5, c8 and c10 are OR-splits and c6, c9 and c11 are OR-joins.

Automatic discovery of care process model is difficult and actual problem. Some methods and algorithms to solve this problem are described in articles [3, 6, 12-16, 19]. Every trace in event logs and in a care

Proceedings of the XVIII International Conference «Data Analytics and Management in Data Intensive Domains» (DAMDID/RCDL'2016), Ershovo, Russia, October 11 - 14, 2016

process model are characterized, in general, the timing associated with the time of applying operations by physician, and is characterized by quality indicators (signs), defining the current state of the patient. In real life we have deviations of real care processes from care plans built on the base of care process model. These deviations can be associated with a change in the patient's state, lack or replacement of some drugs or other medical devices, influence of other disease or other causes.

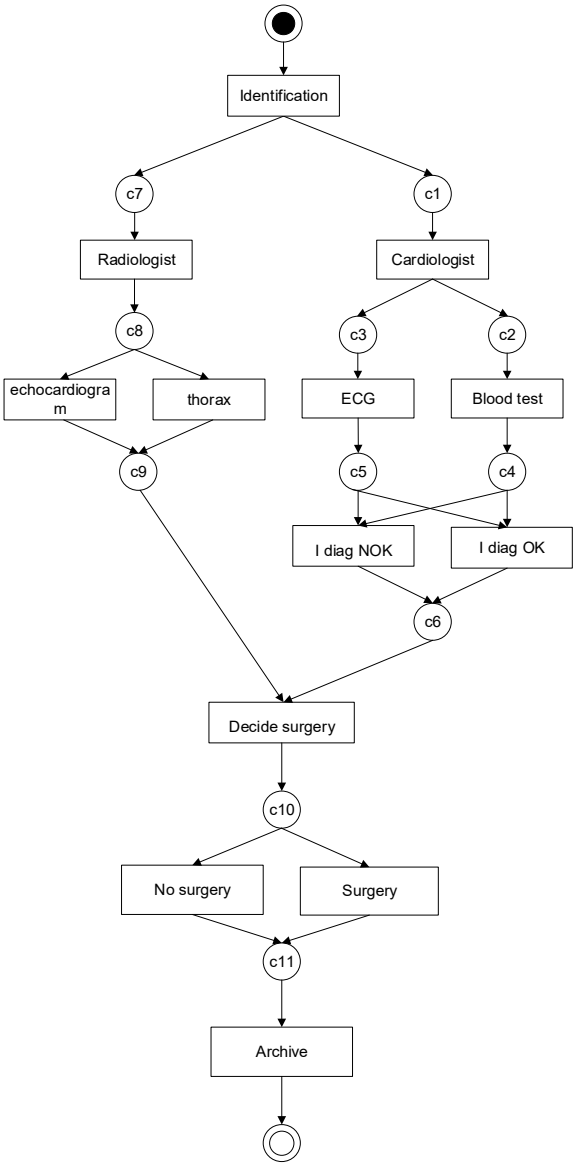


Figure 1 A Petri net of a care process.

3 Identification of possible deviations of a care process from the care plan

Qualitatively expert assessment of a trace as a way on the graph with temporary marks will allow to reveal and estimate various deviations from the course of medical and diagnostic process due to both objective and subjective reasons and to eliminate them in the subsequent realizations of care process. Non-performance of action which has to be surely executed in care process, non-compliance of actions sequence, performance of actions not provided by care process, etc. [2, 9] are considered to be deviations. In care process it is necessary to emphasize the deviations associated with time limits imposed on actions. For example, some actions have to be performed on the first day of the patient arrival. The situation when the action has been performed after the specified time interval is the deviation in this case.

The method for detection and visualization of the deviations associated with performance of actions not included in the model of care process and non-performance of actions that need to be executed is described in the article [10]. In articles [1, 14, 21, 23] the fitness function is used to check conformance of a care process model and processes in event log. Petri nets are used as formal representation of the care process model. Let k is the number of different traces from the aggregated log. For each log trace i, $(1 \le i \le k)$, n_i is the number of process instances combined into the current trace, m_i the number of missing tokens, r_i the number of remaining tokens, c_i the number of consumed tokens, and p_i the number of produced tokens during log replay of the current trace. The token-based fitness metric F is defined as follows [1, 21, 23]:

$$F = \frac{1}{2}\left(1 - \frac{\sum_{i=1}^{k} n_i m_i}{\sum_{i=1}^{k} n_i c_i}\right) + \frac{1}{2}\left(1 - \frac{\sum_{i=1}^{k} n_i r_i}{\sum_{i=1}^{k} n_i p_i}\right)$$

In case of deviation detection in the course of treatment of a specific patient, n and k are equal to 1 since there is one trace and one instance. Function f will be as follows:

$$f = \frac{1}{2}\left(1 - \frac{m}{c}\right) + \frac{1}{2}\left(1 - \frac{r}{p}\right)$$

If f is equal to 1, then the trace is completely consistent with the care process model. Otherwise, there are deviations. In [23] it is shown that deviations from the care plan lead to increased cost of treatment. In [5] the method that could check for deviations of care process of a patient from a plan and locate specific points of these deviations is described.

Let's consider a method for detection of the deviations associated with the performance of actions not provided by the care plan and non-performance of actions provided by care process. The DTW method was applied calculate the deviation or distance. The method allows to find closeness between two measurement sequences for a certain period of time. Generally, the length of sequences can be different, and measurements can be made with different rates [7]. DTW method became widely spread in medicine. A theory of modified DTW algorithms and its applications are presented in [22]. In particular, recognition of human activity is considered by comparison of the gestures presented in

the form of two-dimensional time series. Recognition of such activity of the patients can be very useful in modern healthcare for monitoring of patients condition and automatic reporting creation for health workers. The results of experiments confirmed the sufficient accuracy of one modification.

DTW algorithm calculates the optimal sequence of transformation (deformation) between two time series [7, 20]. Let two numerical sequences (a_1, a_2, …, a_n), (b_1, b_2, …, b_m) are given. We obtain deviations matrix D, where $d_{ij} = |a_i - b_j|$, $i = 1,..,n, j = 1,..,m$. At the second step we build deformations matrix. Each element r_{ij} is defined by means of dynamic programming algorithm and local optimization criteria: $r_{ij} = d_{ij} + \min(r_{i-1, j-1}, r_{i-1, j}, r_{i, j-1})$. The path in the deformation matrix defining a deviation begins in its left upper corner and ends in the right lower. The value R of deformation is defined by the sum of the minimum local deviations of each path element. R is divided by the number of path elements and is considered as distance estimation between sequences.

We consider the example of DTW algorithm application for comparison of processes of bronchial asthma treatment that is presented in Table 1 (data are provided by the Medical center of the Central bank of the Russian Federation). Table contains care plan that includes operations mandatory to perform and reports of real care processes of three patients.

To formalize care process reporting we will define the following variables: «executed» → $\alpha = 1$, «not executed» → $\beta = -1$, «not required» → $\sigma = 0$.The distance between «executed» and «not executed» operations we will define as $|\alpha - \beta| = 2$, and the distance between «executed» («not executed») and «not required», respectively $|\alpha - \sigma| = |\beta - \sigma| = 1$.

Obtained care process parameters are included in Table 2 in the form of sequences

The sequences can now be compared. We will apply the DTW method to calculate care process deviations of care processes from the care plan. In Fig. 2 the process of comparison of the Patient 3 (P3) course of treatment with the plan (P) is presented as comparison of two sequences by the DTW method using the dynamic programming scheme. In the table the way determining the minimum value of deviation R is highlighted in color. In this case R =2.257.

We will apply the DTW method to calculate deviations of Patients 1 and 3 (respectively P1 and P3) care processes. All necessary data for comparison of processes are shown in Fig. 3.

Table 3 demonstrates possibility of distances calculation between precedents. The more the distance, the more precedents are differed from each other. The table shows that Precedent 1 (care trace of the Patient 1) is close to the plan.

Table 1 Bronchial asthma therapy process

Operations of the care plan		Progress report		
		Precedent 1	**Precedent 2**	**Precedent 3**
Reception area / Intensive care unit (ICU)				
1	Transfer from the reception area to the ward with beds /ICU through 2 hours or less	Executed	Executed	Executed
2	External respiration function or Peak expiratory flow rate	Executed	Not required	Not executed
3	Pulse oximetry	Executed	Not required	Executed
4	Chest radiography	Executed	Executed	Executed
1 day in ICU / ward with beds				
5	Pulse oximetry	Not required	Not required	Executed
6	Peak Flow Meter	Executed	Not required	Not executed
7	Inhaled short-acting ß2-agonists or formoterol	Executed	Executed	Executed
2-7 days in ward with beds				
8	Consultation of an exercise therapy doctor	Executed	Executed	Executed
9	Consultation of a physiotherapist	Executed	Executed	Executed
10	Peak Flow Meter	Executed	Executed	Executed
11	External respiration function	Executed	Executed	Not executed
12	Inhaled glucocorticosteroids	Executed	Executed	Executed
13	Inhaled ß2-agonists	Executed	Executed	Executed
8-21 days in ward with beds				
14	Peak Flow Meter	Executed	Not required	Executed
15	External respiration function	Executed	Not required	Executed
16	Systemic glucocorticosteroids	Not required	Not required	Not required
17	Inhaled glucocorticosteroids	Executed	Not required	Executed
18	Inhaled ß2-agonists	Executed	Not required	Executed

Table 2 The summary table of the formalized indicators of care process performance

Transaction number	Precedent 1	Precedent 2	Precedent 3	Care plan
1	1	1	1	1
2	1	0	-1	1
3	1	0	1	1
4	1	1	1	1
5	0	0	1	1
6	1	0	-1	1
7	1	1	1	1
8	1	1	1	1
9	1	1	1	1
10	1	1	1	1
11	1	1	-1	1
12	1	1	1	1
13	1	1	1	1
14	1	0	1	1
15	1	0	1	1
16	0	0	0	1
17	1	0	1	1
18	1	0	1	1

	P3	1	-1	1	1	1	-1	1	1	1	1	-1	1	1	1	1	0	1	1
P		1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18
1	1	0	2	2	2	2	4	4	4	4	4	6	6	6	6	6	7	7	7
1	2	0	2	2	2	2	4	4	4	4	4	6	6	6	6	6	7	7	7
1	3	0	2	2	2	2	4	4	4	4	4	6	6	6	6	6	7	7	7
1	4	0	2	2	2	2	4	4	4	4	4	6	6	6	6	6	7	7	7
1	5	0	2	2	2	2	4	4	4	4	4	6	6	6	6	6	7	7	7
1	6	0	2	2	2	2	4	4	4	4	4	6	6	6	6	6	7	7	7
1	7	0	2	2	2	2	4	4	4	4	4	6	6	6	6	6	7	7	7
1	8	0	2	2	2	2	4	4	4	4	4	6	6	6	6	6	7	7	7
1	9	0	2	2	2	2	4	4	4	4	4	6	6	6	6	6	7	7	7
1	10	0	2	2	2	2	4	4	4	4	4	6	6	6	6	6	7	7	7
1	11	0	2	2	2	2	4	4	4	4	4	6	6	6	6	6	7	7	7
1	12	0	2	2	2	2	4	4	4	4	4	6	6	6	6	6	7	7	7
1	13	0	2	2	2	2	4	4	4	4	4	6	6	6	6	6	7	7	7
1	14	0	2	2	2	2	4	4	4	4	4	6	6	6	6	6	7	7	7
1	15	0	2	2	2	2	4	4	4	4	4	6	6	6	6	6	7	7	7
1	16	0	2	2	2	2	4	4	4	4	4	6	6	6	6	6	7	7	7
1	17	0	2	2	2	2	4	4	4	4	4	6	6	6	6	6	7	7	7
1	18	0	2	2	2	2	4	4	4	4	4	6	6	6	6	6	7	7	7

Figure 2 Calculation of care process deviation from the care plan (on the example of Patient 3)

	P3	1	-1	1	1	1	-1	1	1	1	1	-1	1	1	1	1	0	1	1
P1		1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18
1	1	0	2	2	2	2	4	4	4	4	4	6	6	6	6	6	7	7	7
1	2	0	2	2	2	2	4	4	4	4	4	6	6	6	6	6	7	7	7
1	3	0	2	2	2	2	4	4	4	4	4	6	6	6	6	6	7	7	7
1	4	0	2	2	2	2	4	4	4	4	4	6	6	6	6	6	7	7	7
0	5	1	1	2	3	3	3	4	5	5	5	5	6	7	7	7	6	7	8
1	6	1	3	1	1	1	3	3	3	3	5	5	5	5	5	6	6	6	
1	7	1	3	1	1	1	3	3	3	3	3	5	5	5	5	5	6	6	6
1	8	1	3	1	1	1	3	3	3	3	3	5	5	5	5	5	6	6	6
1	9	1	3	1	1	1	3	3	3	3	3	5	5	5	5	5	6	6	6
1	10	1	3	1	1	1	3	3	3	3	3	5	5	5	5	5	6	6	6
1	11	1	3	1	1	1	3	3	3	3	3	5	5	5	5	5	6	6	6
1	12	1	3	1	1	1	3	3	3	3	3	5	5	5	5	5	6	6	6
1	13	1	3	1	1	1	3	3	3	3	3	5	5	5	5	5	6	6	6
1	14	1	3	1	1	1	3	3	3	3	3	5	5	5	5	5	6	6	6
1	15	1	3	1	1	1	3	3	3	3	3	5	5	5	5	5	6	6	6
0	16	2	2	2	2	2	2	3	4	4	4	4	5	6	6	6	5	6	7
1	17	2	4	2	2	2	4	2	2	2	2	4	4	4	4	4	5	5	5
1	18	2	4	2	2	2	4	2	2	2	2	4	4	4	4	4	5	5	5

Figure 3 Calculation of care processes deviation from the care plan (on the example of Patients 1 and 3)

The distance according to the established scheme is $R = 1.935$. Results of pair comparison of all patients care processes are given in Table 3.

Table 3 Deviations of care processes (pair distances)

	Care plan	Precedent 1	Precedent 2	Precedent 3
Care plan	0,000	0,486	2,143	2.257
Precedent 1	0,486	0,000	0,333	1.935
Precedent 2	2,143	0,333	0,000	4,421
Precedent 3	2.257	1.935	4,421	0,000

3 Conclusion

Medical processes describing care of patients are useful in daily work of physicians, especially in difficult situations. Certain step towards the creation of tools to support physician's work in the course of care process is taken in the article. At the current stage of research, methods for the analysis and evaluation of deviations associated with performance and non-performance of actions for elementary care processes are chosen and studied. The problem is solved on the existing generalized care process model and reduced to evaluation of deviations of care precedent from available traces. The proposed method allows verifying compliance of treatment with the care plan and procedures established by care standard. In addition it helps the decision-maker with the choice of a rational way of the treatment carried out with the use of strategies and rules. In reality, to make a choice of a rational way of treatment it is necessary to sort and estimate rather large number of admissible trajectories taking into account strict binding of operations at the time-point that certainly complicates process of comparison and determination of distances. Authors intend further to research similar processes.

The reported study was funded by RFBR according to the re-search project No 16-37-00034 «Research and development of methods for analysis of deviations of executed medical processes from their model»

References

[1] Wil M. P. van der Aalst, Process Mining: Discovery, Conformance and Enhancement of Business Processes, Springer Publishing Company, Incorporated, 2011, 352 p.

[2] Adriansyah, A.: Aligning Observed and Modeled Behavior. Ph.D. thesis, Eindhoven University of Technology, 2014.

[3] Breiman L., Friedman J. H., Olshen R. A., Stone C. T. Classification and Regression Trees.— Wadsworth, Belmont, California, 1984.

[4] Carroll, Robert J., et al. "Portability of an algorithm to identify rheumatoid arthritis in electronic health records." Journal of the American Medical Informatics Association 19.e1 (2012): e162-e169.

[5] De Leoni, Massimiliano, Fabrizio Maria Maggi, and Wil MP van der Aalst. "Aligning event logs and declarative process models for conformance checking." Business Process Management. Springer Berlin Heidelberg, 2012. 82-97.

[6] Gupta, Shaifali. "Workflow and process mining in healthcare." Master's Thesis, Technische Universiteit Eindhoven (2007).

[7] Ing-Jr Ding, Chih-Ta Yen, Yen-Ming Hsu. Developments of Machine Learning Schemes for Dynamic Time-Wrapping-Based Speech Recognition // Mathematical Problems in Engineering. 2013.

[8] Karthik, R., Menaka, R., Kulkarni, S., & Deshpande, R. (2014). Virtual doctor: an artificial medical diagnostic system based on hard and soft inputs. International Journal of Biomedical Engineering and Technology, 16(4), 329-342.

[9] S. Leemans, D. Fahland, and W.M.P. van der Aalst. Exploring Processes and Deviations. In F. Fournier and J. Mendling, editors, Business Process Management Workshops, International Workshop on Business Process Intelligence (BPI 2014), Lecture Notes in Business Information Processing, Springer-Verlag, Berlin, 2015.

[10] S.J.J. Leemans, D. Fahland, and W.M.P. van der Aalst. Process and Deviation Exploration with Inductive Visual Miner. In L. Limonad and B. Weber, editors, Business Process Management Demo Sessions (BPMD 2014), volume 1295 of CEUR Workshop Proceedings, pages 46-50. CEUR-WS.org, 2014.

[11] Magoulas, George D., Andriana Prentza. "Machine learning in medical applications." Machine Learning and its applications. Springer Berlin Heidelberg, 2001, pp. 300-307.

[12] Mans, Ronny S., Wil MP van der Aalst, and Rob JB Vanwersch. Process Mining in Healthcare: Evaluating and Exploiting Operational Healthcare Processes. Springer International Publishing, 2015.

[13] Mans, R. S., Schonenberg, M. H., Song, M., van der Aalst, W. M., Bakker, P. J. "Application of process mining in healthcare–a case study in a dutch hospital." Biomedical Engineering Systems and Technologies. Springer Berlin Heidelberg, 2008, pp. 425-438.

[14] Mans, R. S., van der Aalst, W. M., Vanwersch, R. J., & Moleman, A. J. "Process mining in healthcare: Data challenges when answering frequently posed questions." Process Support and Knowledge Representation in Health Care. Springer Berlin Heidelberg, 2013, pp. 140-153.

[15] Mans, R. S., Schonenberg, M. H., Song, M., van der Aalst, W. M. P., Bakker, P. J. M. "Process Mining in Healthcare." Case study. Eindhoven University of Technology (2015).

[16] Maruster, L., van der Aalst, W., Weijters, T., van den Bosch, A., Daelemans, W. Automated discovery of workflow models from hospital data. Proceedings BNAIC-01, Amsterdam, October 25-26, 2001, p.183-194.

[17] Nazarenko, G. I., Kleimenova, E. B., Yashina, L. P., Molodchenkov, A. I., Payushchik, S. A., Konstantinova, M. V., Mokin, M.V., Otdelenov, V.A., Sychev, D. A. Development of the ontology of patient management technological records for modeling of clinical workflows in a general hospital. Scientific and Technical Information Processing, 2015, 42(6), pp. 455-462.

[18] Isa, Nor Ashidi Mat. "Towards intelligent diagnostic system employing integration of mathematical and engineering model." AIP Conference Proceedings. Vol. 1660. No. 1. 2015.

[19] Poelmans, Jonas, et al. "Combining business process and data discovery techniques for analyzing and improving integrated care pathways." Advances in Data Mining. Applications and Theoretical Aspects. Springer Berlin Heidelberg, 2010, pp. 505-517.

[20] Romanenko A.A. Time series alignment: forecasting using DTW / Machine learning and data analysis, 2011. Vol. 1, No. 1, pp.77-85.

[21] A. Rozinat and W.M.P. van der Aalst. Conformance Checking of Processes Based on Monitoring 9Real Behavior. Information Systems, 33(1):64–95, 2008.

[22] Mohammad Shokoohi-Yekta, Bing Hu, Hongxia Jin, Jun Wang, Eamonn Keogh. Generalizing Dynamic Time Warping to the Multi-Dimensional Case Requires an Adaptive Approach (SDM 2015). - 2015 SIAM International Conference on Data Mining. - http://www.cs.ucr.edu/~eamonn/Multi-Dimensional_DTW_Journal.pdf (last access: 18.12.2015)

[23] T.J.H. van de Steeg Process Mining in Healthcare: Mining for cost and (near) incidents, Eindhoven, 2015, 67 p.

[24] Taha Samad-Soltani-Heris M. L., Mahmoodvand Z., Zolnoori M. Intelligent Diagnosis of Asthma Using Machine Learning Algorithms. – 2013.

Опыт создания и применения mHealth системы на базе портативного кардиомонитора CardioQVARK

© А.Е.Бекмачев, © С.П.Садовский, © О.В.Сунцова

ООО «Кардиокварк»,
г. Москва
beck@cardioqvark.ru sadovskiy@cardioqvark.ru so@cardioqvark.ru

Аннотация

Статья содержит детальный отчет о создании, принципах функционирования, результатах апробации инновационного телемедицинского комплекса CardioQVARK и возможностях его применения в национальной кардиологической скрининговой системе.

1 Проблема и решение

Статистика неумолима: сердечно-сосудистые заболевания (ССЗ) занимают первое место среди причин смертности населения нашей планеты, опережая птери от эпидемий, войн и изменения климата. По данным Всемирной организации здравоохранения, ежегодная убыль населения в мире по причине ССЗ составляет 17 млн. человек, из них 1,3 млн. - в России [2, 5]. Динамика наблюдений за последние десятилетия эту тенденцию только подтверждает. Особенно прискорбным при нынешнем уровне техники и достижений медицинской науки является то обстоятельство, что у ¾ умерших кардиологические заболевания не были диагностированы при жизни. Этот факт не является специфической особенностью нашей страны, примерно такая же пропорция с различными вариациями наблюдается и в других странах, рис.1.

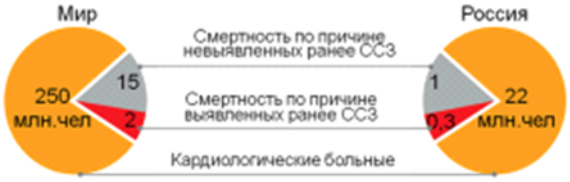


Рисунок 1 Статистика смертности от заболеваний сердечно-сосудистой системы

Очевидным ответом на сложившуюся ситуацию является системная подготовка и проведение комплекса организационно-технических мероприятий по регулярной диагностике максимально широких слоев населения на наличие ССЗ. Такая ранняя диагностика не только в традиционных «группах риска», но и в более обширных социально-возрастных категориях позволит выявлять сердечно-сосудистые и связанные с ними заболевания на ранних стадиях, в более легкой форме, снизить расходы на стационарное лечение, высокотехнологичные медицинские услуги за счет бюджета, уменьшить инвалидизацию населения. Регулярный и системный скрининг также способствуют и развитию психологически позитивной ментальности общества, основанной на здоровом образе жизни.

Уже более ста лет наиболее действенным способом диагностики состояния и здоровья сердца является регистрация биопотенциалов сердца по В.Эйнтховену – запись электрокардиограммы (ЭКГ). Ставшая уже классической методика одинаково хороша как для экспресс-диагностики, так и для системных стационарных исследований. В её основе – регистрация на конечностях и на грудине гальванических токов, вызываемых переменным электрическим полем сердца. Характерные формы сигналов и их комбинации дают информацию об общем статусе сердца и о работе его отделов. Сопоставление этих данных и накопленная за многие десятилетия статистика позволяют проводить диагностику с высокой точностью. До последнего времени электрокардиограф оставался атрибутом медицинского учреждения и даже в портативном варианте, при работе в «полевых условиях», требовал участия квалифицированного медицинского специалиста.

Современные полупроводниковые материалы, технологии микроэлектронки, электронная компонентная база позволили по-новому взглянуть и на такую консервативную отрасль как кардиология и соответствующая медицинская техника. Новейшие технологически достижения и научные разработки дали возможность пересмотреть саму концепцию применения кардиографов, отказавшись от их вечного атрибута – громоздкого кабеля пациента с электродами и фиксаторами, сделать регистратор ЭКГ неимоверно компактным, доступным даже неподготовленным пользователям и таким же привычным, простым в применении, в том числе - и в домашних условиях, как уже ставшие бытовыми электронный термометр или тонометр.

Труды XVIII Международной конференции DAMDID/RCDL'2016 «Аналитика и управление данными в областях с интенсивным использованием данных», Ершово, 11-14 октября 2016

Следующим логичным шагом явилось подключение персонального кардиорегистратора к уже сформировавшейся глобальной информационной структуре «умных» устройств – «интернету вещей» (IoT).

На срезе науки и технологии разработана новейшая российская система CardioQVARK.

2 Принцип работы, состав и характеристики системы CardioQVARK

2.1 Принцип работы

При несомненно инновационном характере, системы CardioQVARK, её идеология и принцип работы предельно просты.

Пациент, имея кардиомонитор, встроенный в чехол мобильного телефона, самостоятельно, в любое время и в любом месте может зарегистрировать ЭКГ, после чего происходит автоматическая отправка данных на «облачный» сервер, который в реальном времени производит обработку информации и возвращает на телефон основные показатели сердечной деятельности, отображаемые в интуитивно понятной текстово-графической форме. Одновременно данные из «облачного» сервера попадают и на планшетный компьютер врача. Приложение для врача имеет расширенный инструментарий для углубленной диагностики кардиограммы, просмотра динамики и встроенные средства прямой обратной связи с пациентом – через электронную почту или службу sms, рис. 2.

Рисунок 2 Принцип и схема работы системы CardioQVARK

2.2 Чехол-кардиомонитор

Ключевой частью рассматриваемой телемедицинской системы является чехол-кардиомонитор. Его революционное отличие – применение жестко встроенных в корпус не металлических, а ёмкостных электродов для съема ЭКГ по I стандартному кардиологическому отведению. В результате стало возможным регистрировать непосредственно переменное электрическое поле сердца, а не индуцируемые им на поверхности тела гальванические токи. Преимущества такого качественного изменения очевидны: отсутствует ненормируемый «вклад» мышечной, сосудистой, кожной проводимости, статического электричества, поляризации электродов, состава физиологических электролитов, пропадает необходимость применения проводящих гелей, теряет значение и сила прижатия электродов к поверхности тела – существенна только полнота перекрытия площади электрода. За счет гальванической развязки по постоянному току при емкостной связи обеспечиваются требования по электробезопасности во время регистрации ЭКГ. Более дружественной стала и сама процедура регистрации кардиосигнала: достаточно запустить приложение на телефоне и приложить к электродам пальцы двух рук, все дальнейшие операции выполнятся в автоматическом режиме.

Благодаря оригинальному схемотехническому решению, верхняя граница полосы пропускания на входе прибора достигла рекордных 10000 Гц (-3 дБ), что по меркам массово применяемых кардиографов с предельным значением полосы пропускания в 150…250 Гц может быть расценено как ЭКГ высокого разрешения.

Источником питания чехла-кардиомонитора служит батарея телефона, с которым прибор соединен через разъем USB, через этот же разъем осуществляется и обмен данными. Потребляемая мощность составляет 17 мВт в режиме ожидания и 90 мВт - в режиме измерения, таким образом, устройство не оказывает существенного влияния на энергетический баланс телефона. Вес прибора не превышает 90 г. Изображения серийных кардиомониторов CardioQVARK в версии для телефонов Apple iPhone 5 / 5s /SE и 6 / 6s приведены на рис. 3.

Рисунок 3 Чехол-кардиомонитор CardioQVARK для телефонов Apple iPhone 5 / 5s / SE и 6 / 6s

2.3 Алгоритмы программного обеспечения

Прогрессивной аппаратной части системы CardioQVARK соответствуют и заложенные в неё алгоритмы. В автоматическом режиме производится расчет, идентификация следующих показателей сердечной деятельности:

Двенадцать основных параметров вариабельности сердечного ритма:

- HR, пульс;
- Extr, экстрасистолия;
- SDNN, стандартное отклонение;
- CV, коэффициент вариации;
- pNN50, разница продолжительности сердечных циклов более 50 мс;

- TP, мощность спектра;
- VLF, очень низкие частоты;
- LF, низкие частоты;
- HF, высокие частоты;
- соотношение LF/HF;
- SI, стресс-индекс;
- ПАРС, показатель активности регуляторных систем.

Реализовано:

- распознавание нарушений ритма сердца;
- построение усредненного кардиоцикла на основе научно обоснованной и клинически апробированной методики [4];
- распознавание импульсов электрокардио-стимулятора на основе собственных прикладных научных исследований [1];.
- распознавание фибрилляции предсердий;
- анализ морфологии кардиоцикла.

3 Апробация

Экспертная оценка врачей по результатам добровольных клинических испытаний показала, что комплексная обработка кардиосигналов в системе CardioQVARK позволяет получать качественную информацию, потенциально пригодную для клинического применения.

С целью документального подтверждения точности получаемых данных в 2015-2016 гг. в семи российских лечебно-профилактических учреждениях (ЛПУ) были проведены добровольные клинические испытания системы. В работе приняли участие: РНЦХ им. Б.В. Петровского (г. Москва), Московский городской научно-практический центр борьбы с туберкулезом (г. Москва), НУЗ «Центральная Клиническая больница № 2 имени Н.А. Семашко» (г. Москва), ГКУЗ МО «ПБ №2 им.В.И. Яковенко» (МО, п. Мещерское), РБУЗ МО «Ивантеевская ЦГБ» (МО, г. Ивантеевка), ПАО «Клиника К+31» (г. Москва), Центр диагностики «Сфера-СМ» (МО, г. Пушкино). Под врачебным контролем также производилась регистрация ЭКГ у пациентов на дому. В испытаниях было задействовано 106 устройств, с их помощью обследовано 876 пациентов. Во время проведения испытаний снято более 12000 кардиограмм. Необходимо отметить, что в ходе проведения добровольных клинических испытаний системы CardioQVARK предотвращено 57 возможных критических случаев. В качестве контрольных применялись кардиографы, уже находящиеся в регулярной клинической эксплуатации, а именно: Kenz Cardioco 1210, Cardio7 Bionet, Biocare ECG-1215, GE MaC 1200 ST, CardioCare, Kenz C1211 и GE MaC 1600.

Наиболее характерные примеры поверки с лентами ЭКГ и комментариями о подтвержденных диагнозах приведены на рис. 4-6.

а)

б)

Рисунок 4 Пациент С., 40 лет. Синдром слабости синусового узла, синкопальные, предсинкопальные состояния, кардиостимулятор Medtronic. а) Kenz Cardico 1210; б) CardioQVARK. Регистрация импульсов кардиостимулятора в реальном времени

а)

б)

Рисунок 5 Пациент Ф., 68 лет. Изменение боковой стенки левого желудочка, постинфарктный кардиосклероз, гипертоническая болезнь. а) MaC 1200 ST; б) CardioQVARK

а)

б)

Рисунок 6 Пациент М., 52 года. Инфаркт миокарда, блокада правой ножки пучка Гиса, желудочковая экстрасистолия. а) CardioCare; б) CardioQVARK

Примеры клинических случаев в том виде, как они отображаются во врачебном ПО CardioQVARK на экране планшетного компьютера iPad, приведены на рис. 7.

а)

б)

в)

г)

Рисунок 7 Клинические примеры в ПО CardioQVARK для врача на экране iPad: а) фибрилляция предсердий; б) постмиокардический кардиосклероз, смена с синусового ритма на ускоренный кардио-вентрикулярный; в) пауза ритма; г) желудочковая экстрасистолия при имплантированном кардиостимуляторе.

В ходе испытаний была подтверждена возможность надежной регистрации кардиомонитором CardioQVARK импульсов имплантированных электрокардиостимуляторов. По результатам испытаний получены отчеты, среди наиболее показательных стоит отметить итоговые документы РНЦХ им. Б.В. Петровского и НУЗ «Центральная Клиническая больница № 2 имени Н.А. Семашко» (г. Москва).

В указанных отчетах подтвержден перечень патологий, которые возможно определить и отследить с использованием системы CardioQVARK. Наименования приведены по Международной классификация болезней 10-го пересмотра (МКБ-10) [3].

Заболевания, которые возможно определить:

- Предсердно-желудочковая (атриовентрикулярная) блокада и блокада левой ножки пучка (Гиса) (I44): I44.0, I44.1, I44.2;
- Пароксизмальная тахикардия (I47): I47.1, I47.2;
- Фибрилляция и трепетание предсердий (I48);
- Другие нарушения сердечного ритма (I49):
- I49.0, I49.1, I49.2, I49.3, I49.4, I49.5.

Заболевания, которые возможно отслеживать в динамике при условии заранее известного диагноза:

- Блокада левой ножки пучка неуточненная (I44.7);
- Другие нарушения проводимости (I45): I45.0, I45.6.

Заболевания, для которых возможно предположить нозологию и выполнить дообследование:

- Ревматически аортальный стеноз с недостаточностью (I06.2);
- Первичная гипертензия (I10);
- Хроническая ишемическая болезнь сердца (I25);
- Перенесенный в прошлом инфаркт миокарда (I25.2);
- Бессимптомная ишемия миокарда (I25.6);
- Стенокардия (I20): I20.0, I20.8, I20.9.

4 Сравнение и анализ

Для оценки уровня техники и рыночных перспектив комплекса было проведено сравнение чехла-кардиомонитора CardioQVARK со сходными по форм-фактору устройствами, доступными на мировом рынке к моменту написания статьи. Ключевым показателем при отборе являлась способность устройства регистрировать ЭКГ по I стандартному отведению, Табл.1.

Анализ показывает, что самым полным набором функциональных возможностей обладает только CardioQVARK.

Важно отметить, что анализ вариабельности сердечного ритма у CardioQVARK реализован строго в системе оценок, рекомендуемых стандартами Европейского кардиологического общества и Североамериканского общества электрофизиологии.

Возможность ведения врачом пациентов с мобильных устройств в режиме online делает комплекс CardioQVARK одним из немногих, соответствующих критерию изделия для мира IoT и отвечающих требованиям применения в среде mHealth.

Часть возможностей CardioQVARK вообще эксклюзивна и не реализована у конкурентов.

Во-первых, это возможность наблюдения за пациентами с имплантированными антиаритмическими устройствами – электрокардиостимуляторами. Такая функция в портативном устройстве реализована впервые и стала возможной благодаря оригинальным решениям в схемотехнике.

Таблица 1 Сравнение CardioQVARK с другими портативными кардиомониторами

Портативные кардиомониторы для регистрации сигналов I стандартного отведения	**CardioQVARK**	**AliveCor**	**ECG Check**	**WMe2**	**Кардиоскоп КС-102**
					
Лента ЭКГ и пульс	●	●	●	●	●
Вариабельность сердечного ритма	●	o	o	●	●
Усредненный кардиоцикл	●	o	o	o	o
Импульсы пейсмейкера	●	o	o	o	o
Фибрилляция предсердий	●	●	●	o	o
Сертификат медицинского изделия / FDA	**август 2016**	●	●	o	o
Программа для врача	●	o	●	o	o
Работа с врачом online	●	o	●	o	o
Облачное хранение	●	●	●	●	o
Открытое API для сторонних разработчиков	●	o	o	o	o

Во-вторых, это открытый интерфейс программирования приложений API, который позволяет сторонним разработчикам разрабатывать своё ПО для взаимодействия с системой CardioQVARK, видоизменять её функционал для своих задач, интегрировать в собственные измерительные и аналитические системы, реализовать кросс-платформенный импорт и экспорт данных.

Полный цикл разработки аппаратной и программной части системы CardioQVARK был проведен в России. Серийное производство чехла-кардиомонитора также налажено целиком на отечественных предприятиях.

Свидетельством высокого научно-технического уровня системы служит обширный перечень объектов интеллектуальной собственности проекта CardioQVARK, включающий в себя программные продукты, полезные модели и промышленные образцы. Количество патентов РФ и заявок на патенты, принятых к рассмотрению в РФ - более десяти, в 2016 г. подана заявка на полезную модель кардиомонитора в системе PCT.

5 Перспективы

Как мы видим, технологии, конструкторские и компоновочные решения, алгоритмы, примененные в составе комплекса CardioQVARK, результаты апробации в клинических условиях позволяют рассматривать его как зрелый аппаратно-программный продукт, пригодный для создания национальной mHealth системы кардиомониторинга и скрининга в реальном масштабе времени.

Простота использования и дружественный интерфейс обеспечивают одинаковую эффективностью применения CardioQVARK на плановом врачебном приеме и при массовой диспансеризации в ЛПУ, контроле нагрузок в физкультурно-оздоровительных и санаторно-курортных учреждениях, а также в бытовых условиях, на дому.

Комплекс с равной эффективностью может применяться в качестве персонального средства оперативного контроля состояния здоровья и пациентами с уже диагностированными кардиологическими заболеваниями, и людьми, ведущими активный образ жизни и проявляющими заботу о своем здоровье еще до возникновения хронических заболеваний и кризисных состояний.

С помощью кардиомонитора возможен сбор однородной обезличенной информации: антропологические показатели пациента, общее состояние здоровья, наличие заболеваний, курс лекарств, ЭКГ-записи с комментариями. Полученные данные доступны для изучения новейшими методами компьютерного анализа и машинного обучения, что в перспективе открывает возможности диагностики заболеваний по ЭКГ и поиску наиболее результативных методов лечения.

Команда проекта CardioQVARK продолжает совершенствовать свой продукт, а имеющиеся наработки и потенциал, заложенный в патентованных решениях, дают основания полагать,

что в скором времени возможности системы выйдут за рамки кардиологии и будут охватывать более широкий спектр задач по персонифицированной диагностике здоровья населения [6].

Литература

[1] Алгоритм выявления мерцательной аритмии в реальном масштабе времени, С.В.Моторина, А.Н.Калиниченко, Журнал «Медицинская техника», 3-2016.

[2] Всемирная организация здравоохранения. Программы и проекты. Сердечно-сосудистые заболевания. http://www.who.int/cardiovascular_diseases/ru

[3] Международная классификация болезней МКБ-10. Электронная версия. http://mkb-10.com/

[4] Построение типового кардиоцикла в системе Cardioqvark, Р.В.Исаков, О.В. Сунцова. Ежегодная Всероссийская научная школа-семинар «Методы компьютерной диагностики в биологии и медицине 2015». Тезисы докладов. Владимирский государственный университет имени А. Г. и Н. Г. Столетовых, 2015 г.

[5] Федеральная служба государственной статистики. Официальная статистика. Население. Здравоохранение. http://www.gks.ru/wps/wcm/connect/rosstat_main/rosstat/ru/statistics/population/healthcare/

[6] Кардиомонитор CardioQVARK. Кардиограмма с помощью телефона http://www.cardioqvark.ru/

Development and application experience of mHealth system based on CardioQVARK portable cardiomonitor

Aleksandr E. Bekmachev, Sergey P. Sadovskiy, Olga V. Suntsova

Details of innovative Russian cardiac monitoring and screening mHealth system CardioQVARK are disclosed. New principles of ECG recording and appropriate application algorithms are discussed. Authors provide the valuable data of product performance validation based on clinical trials. Comparison with the products and solutions of competitors was made and commented. Project perspectives and its commercialization roadmap are indicated.

Построение оптимизированных медицинских конвейерных технологических процессов

© В.М.Хачумов

ФИЦ ИУ РАН,
Москва
vmh48@mail.ru

Аннотация

Описан метод построения конвейерного медицинского технологического процесса. В основе метода лежит процедура оптимизированного закрепления оборудования за операциями и совмещения циклов периодического обслуживания. Приведены примеры построения медицинских конвейеров. Показан формализованный процесс выбора минимально необходимого числа приборов и другой аппаратуры.

1 Введение

Под технологической цепочкой будем понимать конечную последовательность «действий» (операций, этапов, работ, действий и т.д.), направленную на достижение поставленной цели в исследуемой предметной области. Технологический процесс (ТП) содержит, не менее одной технологической цепочки, которые могут быть взаимосвязаны, образуя последовательные, параллельные, параллельно-последовательные и конвейерные конструкции. В дальнейшем полагаем, что технологическая цепочка и технологический процесс являются результатом автоматического планирования (генерации) дискретных «действий», переход между которыми осуществляется в соответствии с набором некоторых правил с учетом темпоральных аспектов. Теоретические вопросы планирования, построения и оптимизации медицинских технологических процессов (МТП) были рассмотрены в трудах отечественных ученых [1,2]. МТП – это система лечебно-диагностических мероприятий, выполнение которых позволяет наиболее рациональным образом провести лечение и обеспечить достижение максимального соответствия научно прогнозируемых результатов реальным. В последнее время большое внимание за рубежом уделяется построению медицинских конвейеров как особых видов МТП. Можно выделить серию работ, посвященную построению различных типов конвейеров, представленных, например, в трудах конференции Pipeline Report (London, 2013) [3,4,5].

Конвейерное обслуживание, например, введено и активно используется в офтальмологии, где построены диагностический и хирургический конвейеры с широким применением современной медицинской техники. Представляет интерес опыт Крымского республиканского центра реабилитации зрения, содержащего лечебно-диагностический офтальмологический конвейер [6], на котором одновременно могут находиться до 100 пациентов, обслуживаемых на специализированных аппаратах, входящих в конвейерный цикл. Для планирования МТП многократного выполнения наилучшим образом подходят периодические расписания с совмещением циклов. Основные положения такой технологии, основанной на эвристических критериях и правилах, изложены в ряде работ [7-10]. Настоящая статья призвана привлечь внимание к возможности построения оптимизированных МТП.

2 Постановка и метод решения задачи

Рассмотрим метод формализации задачи построения периодического расписания с совмещением циклов [7-10]. Пусть последовательность обслуживания производится в соответствии с некоторым технологическим процессом, содержащим m операций, среди которых могут быть одинаковые. Множество разных операций обозначим $\{a_1, a_2, ..., a_l\}$, $1 \le l \le m$, при этом операция a_j в алгоритме выполняется p_j раз, где: $1 \le p_j \le m, \sum_{j=1}^{l} p_j = m.$ В простейшем случае ТП может быть представлен последовательностью операций, в которой результат i-ой операции служит входом для (i+1)-ой операции. Вход первой операции является входом ТП, а результат последней операции - его выходом. Тогда алгоритм можно представить в виде: $A = a_{j_1} a_{j_2} ... a_{j_m}$. Пусть имеется набор Q $\{q_1, q_2, ..., q_l\}$ исполнительных ресурсов l типов, соответствующих операциям МТП. Каждого ресурса q_j j–го типа может быть несколько, т.е.:

Труды XVIII Международной конференции DAMDID/RCDL'2016 «Аналитика и управление данными в областях с интенсивным использованием данных», Ершово, 11-14 октября 2016

$q_j = \{q_{j_1}, q_{j_2}, ..., q_{j_{k_j}}\}$, $\sum_{j=1}^{l} k_j = Q$, $1 \le k_j \le p_j$. ТП назовем отмеченным (ОТП), если в алгоритме A, кроме выполняемых операций, указаны закрепленные за ними ресурсы. ОТП имеет следующий вид: $A^* = a_{j_1}^{i_1} a_{j_2}^{i_2} \ldots a_{j_m}^{i_m}$. Запись $a_{j_k}^{i_k}$ означает, что исполнительный ресурс с номером i_k из набора q_k выполняет операцию a_{j_k} на k-ой фазе реализации ТП. В общем случае одному ТП может соответствовать некоторое множество ОТП. Для организации процесса обслуживания необходимо указать момент начала выполнения процесса. В этом случае при выполнении некоторых дополнительных условий ТП определяет расписание однократного выполнения процесса, показывающее шаг за шагом, какая операция за какой следует и на каком приборе выполняется. Представляет интерес задача многократного выполнения ТП без прерываний. Однократное выполнение ОТП назовем его циклом. Пусть алгоритм выполняется многократно, причем начала циклов совпадают с моментами $\varphi_1, \varphi_2, \ldots$, $\varphi_1 \le \varphi_2 \le \ldots$. При многократной обработке будем оставлять неизменным порядок закрепления приборов, что является одним из необходимых требований обработки без прерываний. Указанное расписание назовем циклическим с периодом T, если для всех его элементов выполняется соотношение $\varphi_j + kT = \varphi_{j+k}$. Будем говорить, что для циклического расписания с периодом T имеет место совмещение вычислительных процессов, если хотя бы для одного элемента расписания выполняется $\varphi_{j+1} < \varphi_j + t_0$, где t_0- время (в тактах), необходимое для выполнения одного цикла. Процесс совмещенной обработки может быть представлен, таким образом, совокупностью взаимодействующих однократных процессов, описываемых ТП. Процесс построения расписания может быть представлен в виде следующей схемы:

1. Генерация вариантов закрепления приборов за фазами выполнения ТП (генерация ОТП).
2. Построение для каждого ОТП периодического расписания с совмещением циклов.
3. Выбор лучшего расписания, минимизирующего среднее времени (в тактах) выполнения одного цикла.

Будем уделять основное внимание вопросам сокращения перебора на каждом из этапов и разработке общей схемы для оперативного планирования периодических расписаний без существенного снижения их качества. Формализуем постановку и решение задачи максимального совмещения циклов выполнения ТП.

Если φ_1 и φ_2- две разные фазы, то величину $\tau = |\varphi_1 - \varphi_2|$ назовем сдвигом фаз. Таким образом, для каждого прибора i-го типа можно выписать множество сдвигов $H_i(A^*)$, а всему варианту поставить в соответствие множество $H(A^*) = \bigcup_{i=1}^{l} H_i(A^*) = \{\tau_1, \ldots, \tau_r\}$. Введем далее множество $G(A^*)$ такое, что $G(A^*) = N_p \setminus H(A^*)$, где $N_p = \{1, \ldots, p+1\}$, $p = \max\{\tau : \tau \in H(A^*)\}$. Множество $H(A^*)$ определяет запрещенные сдвиги между фазами обработки двух соседних партий работ в расписании, а $G(A^*)$ - разрешенные сдвиги. Задача нахождения оптимального варианта ОТП сводится к отысканию оптимального пути на ярусном орграфе (рис.1).

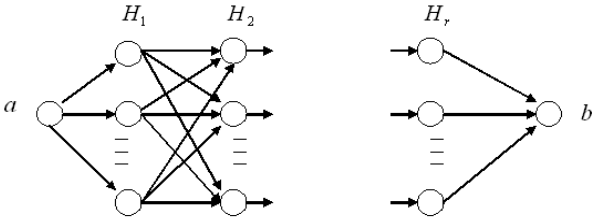


Рисунок 1 Граф задачи

Каждой вершине графа ставится в соответствие один из элементов $H_i, i = 1, \ldots, l.$. Множество несвязанных вершин, отображающее все элементы из H_i, называется i-ым ярусом. Все вершины соседних ярусов соединим дугами, ориентированными от вершин яруса с меньшим номером к вершинам яруса с большим номером. Введем также две дополнительные вершины: начальную a и конечную b, соединив их дугами с вершинами ближайших ярусов. Каждому пути из a в b припишем вес

$r_j(a,b) = |H_j(a,b)|$, $H_j(a,b) = H_{1i_1} \cup \ldots \cup H_{li_l}$;

$j = 1, \ldots, r$; $1 \le i_k \le r_k$. С учетом введенного критерия задача нахождения оптимального варианта ОТП сводится к отысканию такого пути на ярусном орграфе, что $r(a,b) = \min|H_j(a,b)|, j = 1, \ldots, r.$ Т.е. выбирается путь, для которого суммарное число разных сдвигов фаз во множестве $H_j(a,b)$, взятых без повторения, является минимальным.

3 Пример конвейерного МТП

МТП с совмещением циклов позволяют оптимизировать загрузку имеющегося дорогостоящего оборудования (приборов, аппаратов) и минимизировать время обслуживания пациентов врачами. Положения теории совмещения циклов многократно выполняемых процессов рассмотрим применительно построению медицинских глазных хирургических конвейеров. За основу возьмем данные о МТП Республиканского предприятия «Крымский республиканский медицинский центр реабилитации зрения» [6]. Каждому МТП соответствует свой документ – маршрутная процедурная карта аппаратного лечения глаз с соответствующими унифицированными

схемами лечения. Таковыми являются, например, схемы унифицированные СУ-А (табл.1) и СУ-Р (табл.2). Каждая карта содержит описание соответствующего конвейера.

Таблица 1 Маршрутное описание МТП (Унифицированная схема лечения СУ-А)

	Название операции	Обозначение	Приборы, аппараты	Длительность (мин)
1	Биомехано-стимуляция	a	Биомехани-ческий стимулятор "Юность"	5
2	Электро-стимуляция	bbb	Электро-стимулятор "ЭСО-2"	15
3	Лазерплеоптика	cc	Амблиоспекл лазерный АЛ-1	10
4	Засветы панорамными слепящими полями	dd	Панорама	10
5	Упражнения в локализации	e	Амблиотренер	5
6	Нейрогенная релаксация	ffff	Электросон	20
7	Макуло-стимуляция	g	Макуло-стимулятор	5
8	Развивающие программы	hh	Офтальмоло-гический комплекс - МОКК	10

Таблица 2 Маршрутное описание МТП (Унифицированная схема лечения СУ-Р)

	Название операции	Обозначение	Приборы, аппараты	Длитель-ность (мин)
1	Биомехано-стимуляция	a	Биомехани-ческий стимулятор "Юность"	5
2	Электро-стимуляция	bbb	Электро-стимулятор ЭСО-2	15
3	Визотренинг	kk	Визотренер с генератором сигналов	10
4	Оптикофизио-логический массаж аккомодационной мышцы	llll	Аккомодо-тренер АТ-1 Макуло-стимулятор (КЭМБЭЛ)	20
5	Нейрогенная релаксация	ffff	Электросон	20
6	Оптикофизио-логический массаж	pp	Лечебно – тренировоч-ный комплекс ОЛТК-Д	10
7	Компьютерная корректурная программа	hh	Офтальмоло-гический комплекс – МОКК	10

Время заполнения конвейера СУ-А составляет 80 мин, конвейера СУ-Р 90 мин. Как будет показано далее, они могут работать с различными показателями качества в зависимости от наличия оборудования и схемы совмещения циклов. Здесь в качестве основного показателя будем использовать время Т обслуживания конвейером одного пациента после заполнения конвейера.

Рассмотрим решение задачи увязки различных лечебных конвейеров в единую технологическую систему на основе принципов оптимизации совмещения циклов. Каждый из двух представленных лечебных конвейеров представляет МТП, который отличается типом и количеством аппаратов, последовательностью их воздействия на пациента в зависимости от клинических особенностей патологии. Построение общего МТП интересно с точки зрения обобщения операций, реализуемых на одинаковых приборах. Рассмотрим задачу совместного выполнения двух МТП (табл.1 и табл.2). В таблице 3 и на рисунке 2 представлены соответственно варианты оборудования и фрагменты соответствующих диаграмм совмещения циклов обслуживания.

Таблица 3 Влияние объема оборудования на производительность

N	G (A*)	Тип прибора и количество экземпляров											T
		Юность	ЭСО-2	АЛ-1	Панорама	Амблиотренер	Электросон	Макулостимулятор	МОКК	Визотренер	КЭМБЭЛ	ОЛТК-Д	
a	{3}, {5}	1	1	1	1	1	1	1	1	1	1	1	15 25
b	{3}	1	1	1	1	1	2	1	1	1	1	1	15
c	{2}	1	2	1	1	1	2	1	1	1	1	1	10
d	{1}	1	3	2	2	1	4	1	2	1	1	1	5

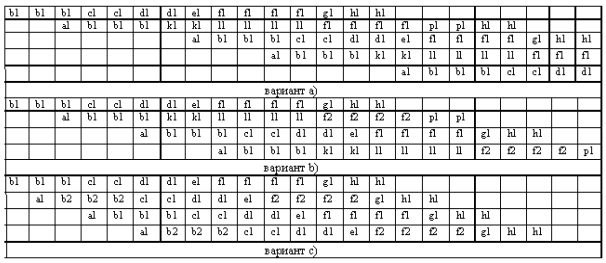


Рисунок 2 Фрагменты диаграмм совмещения циклов работы обобщенного конвейера

1. При минимальном количестве аппаратуры (по одному прибору каждого типа) для двух схем лечения (Вариант a)) конвейер работает неравномерно, обслуживая пациентов с разными заболеваниями через каждые 15 и 25 минут. Интервал между обслуживанием пациентов с одноименными заболеваниями составляет T=40 минут.
2. Вариант b) характеризуется наличием двух приборов типа электросон. Через каждые 15 минут конвейер обслуживает пациентов с чередованием схем лечения. Интервал

между обслуживанием пациентов с одноименной болезнью составит T=30 минут.

3. Вариант c) соответствует равномерному совмещению циклов обслуживания пациентов, при этом T=20 мин.
4. Вариант d) соответствует разделению на два независимых конвейера с достаточным количеством исполнительных ресурсов для достижения максимального быстродействия с показателем T=10 мин.

Если привязать начало работы конвейеров к реальному времени, то помимо указанных показателей из диаграмм совмещения можно определить точное времени работы каждого прибора.

Заключение

Предлагаемый в настоящей работе метод совмещения циклов позволяет оптимизировать повторяющийся многофазовый медицинский технологический процесс при некоторых упрощениях. Лечебные конвейеры обладают возможностями поточного оказания аппаратной и врачебной медицинской помощи населению за счет высокой пропускной способности. Разработка и использование теоретических основ оптимизации медицинских технологических процессов с учетом ограничений в целом позволит: реализовать резервы лечебного процесса в виде материальных ресурсов и времени, выявленных при математическом моделировании; существенно расширить допустимые диапазоны применяемых исполнительных ресурсов, т.е. обеспечить технологическую гибкость. Предлагается в дальнейшем существенно расширить возможности предложенной технологии. Будут предложены правила, позволяющие формировать и оценивать технологические процессы не только состоящие из отдельных цепочек, но и более сложные, содержащие одновременно несколько технологических цепочек, образующие параллельные или параллельно-последовательные схемы. Могут быть сняты некоторые требования на синхронность выполнения операций технологических цепочек. Работа выполнена в рамках проекта РФФИ 16-29-12839 офи_м «Разработка моделей, методов и инструментальных средств для синтеза оптимизированных технологических цепочек и технологических процессов на основе интегрированных баз знаний и интеллектуальных технологий автоматической генерации и оценки планов».

Литература

[1] Назаренко Г.И., Осипов Г.С. Основы теории медицинских технологических процессов. Ч.1. – М.: ФИЗМАТЛИТ, 2005. – 144 с

[2] Назаренко Г.И., Осипов Г.С. Основы теории медицинских технологических процессов. Ч.2. Исследование медицинских технологических процессов на основе интеллектуального анализа. – М.: ФИЗМАТЛИТ, 2006 . – 144 с.

[3] Daniels C. The Tuberculosis Diagnostics Pipeline. – Pipeline Report 2013 – London: 2013, HIV i-base/treatMent aCtIon group, Edited by Andrea Benzacar, pp. 203-221. – http://www.pipelinereport.org/

[4] Lessem E. The Tuberculosis Treatment Pipeline. – Pipeline Report 2013 – London: 2013, HIV i-base/treatMent aCtIon group, Edited by Andrea Benzacar, pp. 223-261. – http://www.pipelinereport.org/

[5] Frick M. The Tuberculosis Vaccine Pipeline. – Pipeline Report 2013 – London: 2013, HIV i-base/treatMent aCtIon group, Edited by Andrea Benzacar, pp. 263-283. – http://www.pipelinereport.org/

[6] Дембский Л.К. Организация реабилитационного «конвейерного» лечения аномалий рефракции, амблиопии, косоглазия в Крымском республиканском центре реабилитации зрения. – http://eyecenter.crimea.com/library/index.html

[7] Хачумов В.М. Модели конвейерного медицинского технологического процесса. – Искусственный интеллект и принятие решений, № 3, 2009, с.25-32

[8] Хачумов В.М. Основные принципы моделирования сложных систем и процессов (Учебное пособие). – М.: Изд-во Российского университета дружбы народов, 2013. – 141 с.

[9] Хачумов В.М. Модель совмещения циклов обслуживания в медицинских технологических процессах. – Materiały ix międzynarodowej naukowi-praktycznej konferencji «wschodnie partnerstwo–2013». – Przemyśl Nauka i studia 2013, pp, 77-80. – http://www.rusnauka.com/26_WP_2013/Informatica/2_144243.doc.htm

[10] Толмачев И.Л., Хачумов М.В. Модели и задачи построения промышленных и медицинских технологических процессов – Приборы и системы, управление, контроль, диагностика. 2013. № 12, с. 38-43.

Construction of optimized medical conveyor processes

Khachumov V.

The method of creation of conveyor medical technological process is described. The method is based on an optimized procedure for fixing of the equipment to operations and combining of periodic service cycles. Examples of creation of medical conveyors are given. The formalized process of selecting the minimum necessary number of devices and other equipment is shown.

Указатель авторов

Научное издание

Аналитика и управление данными в областях с интенсивным использованием данных

XVIII Международная конференция DAMDID / RCDL'2016

Ершово, Московская обл., Россия, 11–14 октября 2016 года

Труды конференции

Data Analytics and Management in Data Intensive Domains

18th International Conference DAMDID / RCDL'2016

October 11–14, 2016, Ershovo, Moscow Region, Russia

Conference Proceedings

Сдано в набор 01.08.16. Подписано в печать 30.09.16.
Формат 60×90/8. Бумага офсетная. Печать цифровая.
Усл.-печ. л. 53,5. Уч. -изд. л. 48,0.
Тираж 120 экз.
Заказ №

Издательство «ТОРУС ПРЕСС»
г. Москва 121614, ул. Крылатская, 29-1-43
E-mail: torus@torus-press.ru
http://www.torus-press.ru

Отпечатано в НИПКЦ «Восход-А»
Москва 109052, ул. Смирновская, д. 25, стр. 3, офис 101
Тел.: 8 (499) 391 34 53, e-mail: admin@vosxod.org
http://www.vosxod.org