Voronezh State University
Federal Research Center
"Computer Science and Control"
of the Russian Academy of Sciences
ACM SIGMOD Chapter

Data Analytics and Management in Data Intensive Domains

Extended Abstracts of the XXII International Conference DAMDID / RCDL'2020

October 13–16, 2020

Voronezh, Russia

Edited by Bernhard Thalheim, Sergey Makhortov,

Alexander Sychev

Voronezh
Publisher "Research publications"
2020

УДК 004.6+004.89
ББК 32.973+32.973.342
Д17

Д17 Data Analytics and Management in Data Intensive Domains: XXII International Conference DAMDID/RCDL' 2020 (October 13–16, 2020, Voronezh, Russia): Extended Abstracts of the Conference. Edited by Bernhard Thalheim, Sergey Makhortov, Alexander Sychev. – Voronezh : Voronezh State University, 2020. – 246 p.

ISBN 978-5-6045486-0-8

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

УДК 004.6+004.89
ББК 32.973+32.973.342

Preface

The XXII International Conference “Data Analytics and Management in Data Intensive Domains” (DAMDID/RCDL’2020) takes place in 2020 on October 13–16 at the Voronezh State University.

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

Traditionally DAMDID/RCDL proceedings are published locally before the conference as a collection of full texts of all contributions accepted by the Program

Committee: regular and short papers, abstracts of posters, and demos.

The program of DAMDID/RCDL 2020 is oriented towards data science and data-intensive analytics as well as on data management topics. The program of this year includes keynotes and invited talks covering a broad range of conference topics.

The keynote by Ladjel Bellatreche (full professor at National Engineering School for Mechanics and Aerotechnics of Aerospace Engineering Group, Poitiers, France) is devoted to models as one of the universal instruments of humans. It aims to motivate researchers to embrace the energy-efficiency of DMSYSs through a survey. A road-map covering the recent hardware and software solutions impacting query processors is proposed and guidelines for developing green query optimizers are widely discussed in the keynote. Oscar Pastor (full professor and director of the Research Center on Software Production Methods at Polytechnic University of Valencia) gives a talk discussing how the use of a sound conceptual modeling support is fundamental in order to convert that "understanding the genome challenge" into a Life Engineering problem, where conceptual modeling, explainable artificial intelligence and data science must work together in order to provide accurate, relia-

ble and valuable solutions, putting an special emphasis in the modern Medicine of Precision applications. Introduction to GaussDB A, an MPP Analytical Database from Huawei is provided by Pavel Velikhov (Principal Software Engineer at Huawei). The research in how to optimize GaussDB A are being conducted in the Moscow Database Lab on this database is briefly described in his invited talk. This research includes using advanced machine learning methods to generate better query plans, solving the problems of skewed data, and incorporating in-memory database technologies. The invited talk by Alexey Molodchenkov (Senior Researcher at Federal research center Computer science and control of RAS) considers an intelligent system for monitoring and controlling factors (health preservation system) that effect the development of diseases. The knowledge base of the system is based on a heterogeneous semantic network, combining the attributes of risk factors into a single network, and includes nodes with recommendations corresponding to individually identified combinations of risk factors.

The conference Program Committee reviewed 60 submissions for the conference and 10 submissions for the PhD workshop. For the conference, 29 submissions were accepted as full papers, 14 as short papers, 3 as a short demo, whereas 14 submissions were rejected. For the PhD workshop, 9 papers were accepted and 1 was rejected.

According to the conference and workshops program, these 55 oral presentations were structured into 11 sessions, including: Data Integration, Conceptual Models and Ontologies; Data Management in Semantic Web; Advanced Data Analysis Methods; Digital Platforms and Information Systems; Data Analysis in Medicine; Data Analysis in Astronomy; Information Extraction from Text.

Though most of the presentations were dedicated to the results of researches conducted in the research organizations located on the territory of Russia, including: Dubna, Ekaterinburg, Innopolis, Kazan, Krasnodar, Moscow, Novosibirsk, Obninsk, Tomsk, Voronezh, Pereslavl, Perm, St. Petersburg, Petrozavodsk, Tver, Tyumen and Yaroslavl, the conference featured international works of talks prepared by the foreign researchers from countries such as Armenia, France, Germany, Spain, United Kingdom, USA.

The chairs of Program Committee express their gratitude to the Program Committee members for carrying out the reviewing of the submissions and selection of the papers for presentation, to the authors of the submissions, as well as to the host organizers from Voronezh State University. The Program Committee appreciates the possibility of using the Conference Management Toolkit (CMT) sponsored by Microsoft Research, which provided great support during various phases of the paper submission and reviewing process.

October 2020

Bernhard Thalheim
Sergey Makhortov
Alexander Sychev

Organization

Program Committee Co-chairs

Bernhard Thalheim	University of Kiel, Germany
Sergey Makhortov	Voronezh State University, Russia
Alexander Sychev	Voronezh State University, Russia

Program Committee Deputy Chair

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

PhD Workshop Chair

Sergey Makhortov	Voronezh State University, Russia

PhD Workshop Curator

Ivan Lukovic	University of Novi Sad, Serbia

Organizing Committee Co-Chairs

Oleg Kozaderov,	Voronezh State University, Russia
Victor Zakharov	Federal Research Center "Computer Science and Control" of RAS, Russia

Organizing Committee Deputy Chairs

Eduard Algazinov	Voronezh State University, Russia
Sergey Stupnikov	Federal Research Center "Computer Science and Control" of RAS, Russia

Organizing Committee

Andrey Koval	Voronezh State University, Russia
Nikolay Skvortsov	Federal Research Center "Computer Science and Control" of RAS, Russia
Dmitry Borisov	Voronezh State University, Russia
Alexey Vakhtin	Voronezh State University, Russia
Dmitry Briukhov	Federal Research Center "Computer Science and Control" of RAS, Russia
Olga Schepkina	Voronezh State University, Russia

Supporters

Voronezh State University

Federal Research Center "Computer Science and Control" of the Russian Academy of Sciences (FRC CSC RAS), Moscow, Russia

Moscow ACM SIGMOD Chapter

Coordinating Committee

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

Nikolay Kolchanov, Institute of Cytology and Genetics, SB RAS, Novosibirsk, Russia (Co-Chair)

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

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
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

Program Committee

Ladjel Bellatreche	Laboratory of Computer Science and Automatic Control for Systems, National Engineering School for Mechanics and Aerotechnics, Poitiers, France
Dmitry Borisenkov	Relex company, Russia
Pavel Braslavski	Ural Federal University, Russia
Vasily Bunakov	Science and Technology Facilities Council, Harwell, UK
George Chernishev	Saint-Petersburg State University Russia, Russia
Boris Dobrov	Research Computing Center of Lomonosov Moscow State University, Russia
Alexander Elizarov	Kazan Federal University, Russia
Alexander Fazliev	Institute of Atmospheric Optics, SB RAS, Russia
Yuriy Gapanyuk	Bauman Moscow State Technical University, Russia
Veronika Garshina	Voronezh State University, Russia
Evgeny Gordov	Institute of Monitoring of Climatic and Ecological Systems SB RAS, Russia
Valeriya Gribova	Institute of Automation and Control Processes FEBRAS, Far Eastern Federal University, Russia
Maxim Gubin	Google Inc., USA
Sergio Ilarri	University of Zaragoza, Spain
Mirjana Ivanovic	University of Novi Sad, Serbia
Nadezhda Kiselyova	IMET RAS, Russia
Vladimir Korenkov	Joint Institute for Nuclear Research, Russia
Sergey Kuznetsov	Institute for System Programming, RAS, Russia
Evgeny Lipachev	Kazan Federal University, Russia
Natalia Loukachevitch	Lomonosov Moscow State University, Russia
Ivan Lukovic	University of Novi Sad, Serbia
Oleg Malkov	Institute of Astronomy, RAS, Russia
Yannis Manolopoulos	School of Informatics of the Aristotle University of Thessaloniki, Greece
Archil Maysuradze	Lomonosov Moscow State University, Russia
Manuel Mazzara	Innopolis University, Russia

Alexey Mitsyuk	National Research University Higher School of Economics, Russia
Xenia Naidenova	Kirov Military Medical Academy, Russia
Dmitry Namiot	Lomonosov Moscow State University, Russia
Dmitry Nikitenko	Lomonosov Moscow State University, Russia
Panos Pardalos	Department of Industrial and Systems Engineering, University of Florida, USA
Natalya Ponomareva	Research Center of Neurology, Russia
Alexey Pozanenko	Space Research Institute, RAS, Russia
Roman Samarev	Bauman Moscow State Technical University, Russia
Timos Sellis	Swinburne University of Technology
Vladimir Serebryakov	Computing Centre of RAS, Russia
Nikolay Skvortsov	Federal Research Center "Computer Science and Control" of RAS, Russia, Russia
Manfred Sneps-Sneppe	AbavaNet
Sergey Sobolev	Lomonosov Moscow State University, Russia
Valery Sokolov	Yaroslavl State University, Russia
Alexey Ushakov	University of California, Santa Barbara, USA
Pavel Velikhov	Huawei
Alexey Vovchenko	Federal Research Center "Computer Science and Control" of RAS, Russia, Russia
Vladimir Zadorozhny	University of Pittsburgh, USA
Yury Zagorulko	Institute of Informatics Systems, SB RAS, Russia
Victor Zakharov	Federal Research Center "Computer Science and Control" of RAS, Russia, Russia
Sergey Znamensky	Institute of Program Systems, RAS, Russia
Mikhail Zymbler	South Ural State University, Russia

Abbreviations

EPFL	École Polytechnique Fédérale de Lausanne
FRCCSC RAS	Federal Research Center "Computer Science and Control" of the Russian Academy of Sciences
IMET RAS	A. A. Baikov Institute of Metallurgy and Materials Science of Russian Academy of Sciences
IMM UB RAS	N. N. Krasovskii Institute of Mathematics and Mechanics of the Ural Branch of the Russian Academy of Sciences
INASAN	Institute of Astronomy of the Russian Academy of Sciences
ISTP	Institute of Solar Terrestrial Physics, Irkutsk
ITEP	Institute of Theoretical and Experimental Physics
JIHT RAS	Joint Institute for High Temperatures of the Russian Academy of Sciences
KFU	Kazan Federal University
KIAM PSU	Keldysh Institute of Applied Mathematics, Petrozavodsk State University
MIPT	Moscow Institute of Physics and Technology
NRU HSE	National Research University Higher School of Economics
RUDN	Peoples' Friendship University of Russia
SAAO	South African Astronomical Observatory, Cape Town
SAI MSU	Lomonosov Moscow State University, Sternberg Astronomical Institute
SB RAS	Siberian Branch of the Russian Academy of Sciences
SRI RAS	Space Research Institute of the Russian Academy of Sciences
STFC	Science and Technology Facilities Council, UK
VSU	Voronezh State University

Contents

Data Management in Semantic Web

Advanced Data Analysis Methods

Data Analysis in Medicine

Data Analysis in Astronomy and Spectral Data

Information Extraction from Text II

PhD Workshop

KEYNOTES AND INVITED TALKS

Towards Green Data Management Systems (Extended Abstract)

Ladjel Bellatreche [0000-0001-9968-0066]

National Engineering School for Mechanics and Aerotechnics of Aerospace Engineering Group, Poitiers, France
ladjel.bellatreche@ensma.fr

In today's world, our life depends too much on computers. Therefore, we are forced to look at every way to save the energy of our hardware components, system software, as well as applications. Data Management Systems (DMSYSs) are at the heart of the energy new world order. The query processor is one of the DMSYS components in charge of the efficient processing of data. Studying the Energy-Efficiency of this component has become an urgent necessity. Most query optimizers minimize inputs/outputs operations and try to exploit RAM as much as possible. Unfortunately, they generally ignore energy aspects. Furthermore, many researchers have the opinion that only the OS and firmware that should manage energy, leaving DMSYSs as a second priority. In our opinion, software and hardware solutions must be integrated to maximize energy savings. This integration seems natural since query optimizers use cost models to select the best query plans and use hardware and software parameters. As scientists, we first feel obliged to motivate researchers to embrace the Energy-Efficiency of DMSYSs through a survey. Secondly, to accompany them, we propose a road-map covering the recent hardware and software solutions impacting query processors. Finally, guidelines for developing green query optimizers are widely discussed.

Conceptual Modeling and Life Engineering: Facing Data Intensive Domains Under a Common Perspective (Extended Abstract)

Oscar Pastor [0000-0002-1320-8471]

Polytechnic University of Valencia, Research Center on Software Production Methods
opastor@dsic.upv.es

Understanding the Hunan Genome is the big scientific challenge of our century. It will probably lead to a new kind of "Homo Sapiens" with new capabilities never before affordable for human being as we know them. This will be referred in the keynote as an "Homo Genius" evolution. Getting a shared understanding of the domain is a first essential task in order to manage correctly and efficiently such a complex data intensive domain. With more and more data being generated day after day with the continuous improvements of sequencing technologies, selecting the right data management strategy intended to support the design of the right software platforms that successfully attend user requirements (in terms of relevant information needs) becomes a unavoidable, crucial goal. This talk will discuss how the use of a sound conceptual modeling support is fundamental in order to convert that "understanding the genome challenge" into a Life Engineering problem, where conceptual modeling, explainable artificial intelligence and data science must work together in order to provide accurate, reliable and valuable solutions, putting an special emphasis in the modern Medicine of Precision applications.

Huawei GaussDB A and Moscow Database Intelligence and Optimization Technology Center (Extended Abstract)

Pavel Velikhov [0000-0002-0644-8047]

Huawei
pavel.velikhov@gmail.com

In this talk GaussDB A, an MPP Analytical Database from Huawei, is introduced. The research are being conducted in the Moscow Database Lab on this database is briefly described. GaussDB A is a world-class Analytical Database with many deployments in mainland China and over the world. It has been in development for over 8 years now. Currently, its being offered as a standalone solution and also as part of Huawei Cloud. In the Moscow Database Lab research in how to optimize GaussDB A, including using advanced machine learning methods to generate better query plans, solving the problems of skewed data, and incorporating in-memory database technologies are being conducted.

Artificial Intelligence and Data Analysis Methods in Healthcare (Extended Abstract)

Alexey Molodchenkov [0000-0003-0039-943X]

Federal Research Center "Computer Science and Control",
Russian Academy of Sciences, Moscow, Russia
molodchenkov-ai@rudn.ru

Здоровье определяется многочисленными факторами: наследственными, средовыми, социальными и др. Их совокупность определяет риск возникновения различных патологий. В ФИЦ ИУ РАН ведется разработка технологии оценки рисков различных заболеваний и построения плана профилактических мероприятий с использованием методов искусственного интеллекта и анализа данных. При этом здоровьесбережение и оздоровление организма рассматривается в цифровом контуре профилактической медицины. В связи с этим следует отметить, что профилактика в традиционной форме, то есть, как комплекс мероприятий, направленных на предупреждение какого-либо явления и/или устранение факторов риска, приводящих к этому явлению, включает формирование здорового образа жизни, в том числе повышение уровня знаний всех категорий населения о влиянии негативных факторов и возможностях снижения этого влияния. Разработанная в ФИЦ ИУ РАН интеллектуальная система для мониторинга и контроля факторов, угрожающих развитием заболеваний, включает формирование индивидуальных мероприятий по оздоровлению людей у которых обнаружены риски отдельных, социально значимых заболеваний (инфаркт, инсульт, депрессия). В основе рассмотренной системы здоровьесбережения заложен принцип медицины 4P: предсказание (predictive) обеспечивает прогноз или оценку риска заболеваний, предупредительные мероприятия (preventive) позволяет формировать рекомендации, основанные на учете личностных (personalize) особенностей проявления возможной патологии, а соучастие (participant) предполагает активное участие человека в коррекции своего образа жизни. Эти принципы 4P обеспечиваются благодаря тому, что база знаний системы на основе неоднородной семантической сети не только объединяет атрибуты факторов риска в единую сеть, но и включает туда узлы с рекомендациями, соответствующими выявляемым в конкретный момент комбинациям риск-факторов у индивидуума. Построение персональных рекомендаций основано на алгоритме аргументационных рассуждений. Это соответствует персонализированному подходу к профилактике болезней. Поступающие в базу данных новые сведения позволяют отслеживать и учитывать критически важные отклонения параметров в динамике. Сама база знаний и ряд интеллектуальных компонентов системы были построены на основе анализа большого объема данных.

DATA INTEGRATION,
CONCEPTUAL MODELS AND ONTOLOGIES

Intelligent Systems Multidimensional Architecture Conceptual Modeling (Extended Abstract)

Konstantin Kostenko [0000-0002-9851-2455]

Kuban State University, Stavropolskaya str. 149, 350040, Krasnodar, Russia
kostenko@kubsu.ru

The intelligent systems unified architecture proposed for comprehensive modelling the subject domains specialists' activities in professional tasks solving. The human thinking processes and memory structures invariants adapted to given activity areas form foundations for such modelling.

General intelligent system concept suppose applying the complete set of independent knowledge attributes considered as having key priority at science areas concerned on studies the knowledge concept. Separate dimension's meanings form ordered set of knowledge attributes possible meanings distributed through levels, associated with knowledge quants sets. Quants adopted as knowledge fundamental parameters meanings that define formalized knowledge dimensions. The structural and functional aspects of proposed architecture developed by analyzing and integrating the entities of different models in mathematics, system engineering, linguistics and cognitive science.

Such architecture components formalized specifications make a basis for creating the uniform descriptions of formalized cross-disciplinary intelligent systems models. They allow further transforming into intelligent systems' applied prototypes with unified common structural and functional properties. These prototypes based on the principles of knowledge engineering and able to accumulate specialist's experience in keeping and processing the complex knowledge structures.

The two-dimension intelligent system architecture investigated. It relates to aspects of knowledge representation abstractness and atomization. The aspect of knowledge abstractness is similar to K. Stanovich idea of multilevel architecture for one-dimension intelligent system with external, algorithmic and abstract quants for knowledge abstractness dimension.

External memory level proposed for modeling the processes of interaction with external entities of knowledge area's content representation. Such content exists outside the intelligent system and presented by knowledge-contained resources.

The algorithmic memory level's application relates to modeling professional tasks' algorithmic solving by intelligent agents associated with this level's components.

Components of abstract memory supply intelligent system with entities that define abstract invariants for modelling knowledge representation formats and morphisms for simulating the intelligent system functional aspects. Mathematical invariants' coordination with cognitive aspects of knowledge processing modeling based on classifiers for goals and operations as thinking processes' elements. The intuitively complete basic system of cognitive goals introduced by B. Bloom and revised by D. Krathwohl.

The structure's second dimension quants correspond to knowledge representations decomposition degree meanings. Knowledge quants define following considered dimension's meanings: completely atomized (expressed by ontology representation close to descriptive logics' formats), partially broken (intermediate formats with knowledge aggregates, useful for subsequent assembling into complex knowledge images) and full images (complete representations of subject area content fragments realizations).

Formal basis for unified intelligent system architecture and knowledge representation formats depends on operations, used at abstract mathematical models. Operations' unified format suppose that they are unary with only exception of binary composition operation. Last operation allows integrating different objects into their compositions (direct sums) used as knowledge representation formalisms' invariant. Computable algebraic operations with given domains and ranges simulate knowledge transformations. They based on unified content's representation formats for intelligent system components' memory. Operations are integrated at classes and analogous to wide spectrum of abstract operations proposed within fundamental mathematical models. They have exact semantics based on operations' formal descriptions.

General formats for knowledge flows and processes descriptions define them as knowledge transforming within and transferring between architecture's components. Such flows structure demonstrates knowledge processing stages as its transformation or transferring by transition between components that are neighbors in one dimension.

Scenarios look like oriented graphs (diagrams) with vertices, marked by names of operations classes. Abstract scenario formal description realize proposition that one operations class originates elementary scenario, formed by one diagram vertex marked with class name. Complex scenario looks as combination of elementary scenarios, connected by directed edges.

Abstract scenario's diagram presents knowledge-based process' initial description that allows transforming into diagram's homomorphic extensions. Such transformations performed in several steps. They lead to gradual extending process's description by parameters' additional specifications. The diagram's homomorphic extensions implemented by operations of process model parameters' adding, splitting and restriction. These operations allow diagrams' transforming into their detailed descriptions with last diagrams' inverse transformation into previous diagrams by diagrams' homomorphisms. The first operation introduces into processing the new attributes that extend the diagrams' and diagram elements structure with additional parts. This allows new parameter embedding into diagrams' descriptions. Second operation relates to diagram's existing attributes. They structured by splitting on components with possibility of these components' values transforming backward into whole ones. The parameters' restriction operations served as tool for attributes' values domains narrowing for initial diagrams' transformation into their more exact variants. All these operations allow inversions and their compositions define diagrams homomorphisms by compositions of homomorphic extensions' inversions. Operations combinations define the operations' complex descriptions for goals' realization diagrams. Operations' general descriptions identify operations' properties and presented by formal expressions.

Diagrams constructed of two kinds of elements (vertices and directed arrows). In such format they seemed too abstract and weakly compliant with cognitive processes reality. The abstract diagram's transforming into subject area task solving knowledge-

based process may demand performing a long sequence of extensions. They perform diagram's direct transforming into such one that sets exact knowledge processing description and has algorithmic realization. The first diagram as abstract diagram. It presented by graph with vertices marked by operations classes' names. Such diagram extended by operations' domains and ranges vertices. New diagram presents the initial diagram's homomorphic extension. Such diagram's format based on initial diagram extension by adding new vertices. They change the vertices' selected properties by pointing out the domains and ranges for classes of operations assigned to initial diagram's vertices. The last diagram operation demonstrates diagram developing by adding conditional expressions associated with diagram's arrows. Conditions inserted into diagrams allow describing different ways of diagrams' applications when solving subject area tasks. Separate conditional expression allows inclusion into diagram with additional attribute of its role. This attribute specifies way of condition interpreting by intelligent systems' processes modeling algorithms.

Templates form basic description for realizing knowledge area goals by knowledge flows and processing between and within intelligent system components. Thoroughly developed templates and diagrams may settle the applied intelligent system complete model. The diagrams' based knowledge processing uses diagrams' formal descriptions and operations' implementation algorithms. The knowledge flow element demonstrates knowledge transition between two neighbor intelligent system's components. The corresponding knowledge flow formal descriptions based on these components' common knowledge format. The knowledge for intercomponent transferring allocated within flow's initial component memory. Appropriate knowledge format defines its structure.

Knowledge transferring uses knowledge structure morphism, based on saving the initial knowledge algebraic structure and replacement structures' elements with elements adopted at transferring end components.

The developed models allow consider the thinking processes invariants as a basis for technology of designing the intelligent systems at specialists' activity areas. The proposed intelligent systems developed as results of adaptation the thinking processes' and memory structures' universal invariants. They allow performing complete simulating the subject specialists' professional activity. Mathematical basis for described intelligent systems' models originates from abstract concept of knowledge representation formalism. These formalisms' invariants, used as basic for formal specifications of nonmathematical concepts, essential for representation the human memory structures and thinking processes. Intelligent systems' universal components form uniform structure. The last one based on knowledge dimension concept and realized by knowledge flows that cross separate dimensions and performed within such a structure.

The subject area's knowledge flows and processes concept integrates entities developed within a great number of knowledge areas that deal with human intelligence modeling. These systems' formal integration is possible by following the cybernetic principles as transdisciplinary approach to modeling living, social and technological systems by goals' driven information flows. The multi-dimension intelligent systems architecture allows such systems designing as structures created from universal unified components. Homomorphic extensions form base for templates and diagrams'

multilevel modeling. *Time* and *grade* knowledge aspects extend number of intelligent systems' dimensions.

The reported study was funded by RFBR and administration of Krasnodar territory grant project number № 19-41-230008 and by RFBR grant project number № 20-01-00289.

Managing Data-Intensive Research Problem-Solving Lifecycle (Extended Abstract)

Nikolay A. Skvortsov and Sergey A. Stupnikov

Institute of Informatics Problems,
Federal Research Center "Computer Science and Control",
Russian Academy of Sciences, Moscow, Russia
nskv@mail.ru

Data management to ensure their reuse and reproducibility of research results has been a pressing issue for many years. An excellent summary of previous efforts is the FAIR data principles [5], which declare forward-looking requirements for data management. Supporting data with rich metadata, provenance information, providing access, searching, and using semantic technologies are prerequisites for achieving data interoperability and reuse. The FAIR data principles have become one of the main directions for actively discussed foundations for creating global interdisciplinary research data infrastructures. There are communities of researchers interested in achieving these goals working with data infrastructures in their disciplines. In this context, the purposes of the discussion in the paper are possible problem-solving lifecycle in data infrastructures and proposing the semantic specifications of domains as its core.

In [3] a lifecycle of research was proposed, which was considered as a reflection of the FAIR data principles. The heart of this lifecycle is formal domain specifications maintained by communities working in those domains. Management of domain descriptions makes possible searching for data relevant to the research and for methods to reuse them, then solving problems in terms of domain specifications and represent research results in the same terms. In this way, research results remain reachable for reuse in other research within the community.

This paper proposes an enhancement of the lifecycle of research problem solving which can combine both simplified accustomed approaches of research groups to data analysis, and the possibility of semantic approaches to managing heterogeneous data and formal verification of their correct use at different stages of integration in solving problems.

Most disciplines have become data-intensive domains and use data resources to discover new knowledge from them. Different kinds of research meet certain difficulties with data management. It is necessary to find relevant data, work with heterogeneous data resources of various provenance. The research process is very sensitive to the needs for semantic integration of data and to the availability of implemented methods. Researchers spend the most effort on these needs and are interested in the reuse of correct data and method integration.

There has always been a need to tie data with methods that can process it or define the inherent behavior of entities described by data. Libraries of methods related to a domain suffer from the ability to include only basic methods, it is not possible to extend them with the results of various solved research problems.

During problem-solving using multiple data resources, researchers meet data heterogeneity problems related to the different understanding of the subject area concepts (ontological level), representation of data using different languages or types of databases (data model level), different data structure and types (conceptual schema level), and presence of different data about the same entities in different data sources (object identity level). Resolving data heterogeneity sometimes takes a lot of effort and time during research. It often becomes a cause of incorrect interpretations of data and errors in research results. Therefore, the problem of heterogeneity of data should be paid the closest attention in research data infrastructures. It is preferable to develop thorough solutions to these problems not repeatedly but based on reuse of the made decisions.

To provide the process of research problem-solving in data infrastructures, on the one hand, it is necessary to have a set of services to develop a problem statement, search for data resources and methods, integrate them at all the levels described above. On the other hand, having good opportunities for resource integration, it is necessary to minimize using them and maximize their reuse in many research problems being solved.

It becomes possible when a domain community uses and supports domain models and research problems are formulated using them. Resources containing data and methods could be integrated into the same domain. In this paper, domain models are called specifications since they represent the basic requirements in terms of which research is conducted in the community. They use specification languages and they are convenient to formulate exact problem specifications based on them. The idea is not about globally centralized data representations. There are different data representations supported and traditionally used in communities, but findable mappings between them could be developed if necessary. This does not restrict researchers from being able to express special things in their research. However, common parts are standardized, and it ensures interoperability in communities. More specific narrow communities can describe their subdomains using the specifications of broader communities. Opposing communities may appear, but their representations can be integrated if necessary.

Wide communities and thematic groups are interested in the development of domain specifications for research data infrastructures. And both resource creators and users are interested in the integration of resources in such specifications. In this sense, research communities in data infrastructures should play a major role in maintaining domain specifications.

First of all, metadata registry services are introduced, which form the basis for the interaction of all stages and activities in problem-solving. Registries are intended for different kinds of specifications and corresponding collections of resources in data infrastructures. The following registries are proposed:

- domain specification registry containing formal ontologies of research domains, special ontologies like the provenance ontology [2], and conceptual schemas linked to the ontologies;
- data model registry for describing integrated data model elements;
- resource registry for the integration of data sources relevant to domains;
- method registry for specifying implementations of methods and processes applicable to the domain objects;
- digital object registry is used to support current research projects, requirement specifications over the domain specifications; links to integrated data resources and methods used in problem-solving; implemented programs; intermediate results.

The baseline lifecycle of research problem-solving includes problem specification, data resource selection, method selection, and data processing for problem-solving itself. However, each of these stages can be accompanied by a set of services to achieve interoperability at all levels and reuse data, methods, and integration results. Any activities in the problem-solving lifecycle focus on the analysis of metadata stored in registries and reflect the results in them. The problem-solving lifecycle can include the following stages.

Problem statement specifications are developed as a specialization of domain specifications. It normally begins with the specification of the requirements, ontological specifications are developed, and the problem is specified in terms of conceptual schemas. Semantic annotations on the ontological level link the requirements, the conceptual schema elements, and later used for discovering and mapping different types of resources that can potentially be used for solving it.

Data resource selection and reuse can be supported by a comprehensive set of means that provide a semantic search for relevant data resources in the resource registry and their semantic integration including data model integration, schema mapping, integration at the object level, and formal verification of integration correctness.

Data model unification [4] creates a set of extensions of the canonical model in the data model registry that maps elements of resource data models expressing them with specific structures, data types, operations, or constraints in the canonical model.

Then the conceptual schemas of data resources represented in a unified data model are integrated. Data resource schemas are preliminarily linked by ontological descriptions and mapped to the data representations accepted in the domain. Structural conflicts of schemas are reconciled.

After schema mapping and implementing requirement entity types with a set of relevant data resources, it is necessary to identify the same real-world entities among data from different resources, and correctly compile objects of the domain entity types. Object identification criteria are developed and entity resolution is performed in the domain independently of specific data resources. Data fusion rules for identified entities may depend on the problem being solved.

After integrating data resources, method selection can be started with a search in the method registry. Some implementations of methods related to specific research objects may be associated with data resources, some are separate resources such as

services and libraries. Existing implementations of relevant methods should be selected and integrated, or new methods are developed to solve the problem.

Formal verification of specification mapping on every stage is preferable. Specifications are formalized in the abstract machine notation (AMN), and the B-technology [1] is used for semi-automated prove of the refinement relation between mapped specifications.

To maintain shared specifications in the research community, the final and key stage of problem-solving is publishing to metadata registries. It reflects there not only the resulting data but also the results of all stages of problem-solving:

- new domain knowledge (ontologies) approved by the community;
- additions to conceptual schemas of the domain approved by the community;
- extensions of the canonical data models, results of the mapping of heterogeneous data models to them;
- available data resources and results of integration including their mappings to the conceptual schemas of the domain and ontological descriptions of their elements to provide the search for resources;
- implementations of methods and processes allowing the reproducibility of results with ontological descriptions of their semantics, semantics of their input and output parameters;
- research results data including new data being the goal of problem-solving, source data gathered from multiple sources, converted to a unified representation, selected with certain criteria, related to certain entity types, enriched data, and others, as well as the metadata related to their schema mapping, semantic annotations, and provenance.

The described problem-solving lifecycle allows to draw the conclusions about the necessary services supporting it in data infrastructures to implement the FAIR data principles: They can include registries, metadata publishing and maintenance services, semantic search services using ontologies and published semantic annotations, data integration services including integration of their data models, schema mapping, entity resolution, data fusion, method and process integration, and finally formal semantics verification services ensuring the correctness of specification mappings on different levels.

References:

1. Abrial, J.-R.: The B-Book: Assigning Programs to Meanings. Cambridge: Cambridge University Press (1996)
2. Belhajjame K., Cheney J., Corsar D., Garijo D., Soiland-Reyes S., Zednik S., Zhao J. PROV-O: The PROV Ontology. W3C Recommendation. W3C (2013). https://www.w3.org/TR/prov-o.
3. Skvortsov, N A. Meaningful data interoperability and reuse among heterogeneous scientific communities. In: Selected Papers of the XX International Conference on Data Analytics and Management in Data Intensive Domains (DAMDID/RCDL 2018). L. Kalinichen-

ko,Y. Manolopoulos, S. Stupnikov, N. Skvortsov, V. Sukhomlin (Eds.), Vol. 2277, P. 14-15, CEUR (2018). http://ceur-ws.org/Vol-2277/paper05.pdf.

4. Stupnikov S., Kalinichenko L. (2019) Extensible Unifying Data Model Design for Data Integration in FAIR Data Infrastructures. In: Manolopoulos Y., Stupnikov S. (eds) Data Analytics and Management in Data Intensive Domains. DAMDID/RCDL 2018. Communications in Computer and Information Science, vol 1003, P. 17-36. Springer, Cham. doi.org/10.1007/978-3-030-23584-0_2
5. Wilkinson M., et al. The FAIR Guiding Principles for scientific data management and stewardship. In: Scientific data, Vol. 3 (2016).

Algebraic Models for Big Data and Knowledge Management (Extended Abstract)

Sergey Makhortov [0000-0001-6362-4468]

Voronezh State University, 1, Universitetskaya pl., Voronezh, 394018, Russia
msd_exp@outlook.com

Abstract. Effective formalism for the construction and study of data and knowledge models is provided by algebraic methods. In particular, the formal methodology for knowledge management in production-type systems is developed by the lattice-based algebraic theory of LP-structures (lattice production structures). Another achievement of the theory is the method of relevant backward inference (LP-inference), which significantly reduces the number of queries to external sources of information. This result is especially relevant when processing big data and managing large volumes of knowledge. The basis of LP-inference is the apparatus of production-logical equations. The report for the first time introduces an extended class of such equations for an algebraic model, the expressive capabilities of which cover logical inference systems with a fuzzy knowledge base. Thus, the advantages of the theory of LP-structures extend to a fuzzy logical inference.

Keywords: fuzzy production system, relevant backward inference, algebraic model, fuzzy LP-structure, production-logical equation.

1 Introduction

When processing large volumes of information, obtaining an accurate solution often requires excessive resources or is even impossible at all. Therefore, in modern information systems, methods of artificial intelligence and inference are used, which allow to find approximate solutions with a given accuracy in an acceptable time. An important feature of such systems is represented by the fuzzy nature of knowledge and reasoning [1]. The basis for their construction can be knowledge bases with some simplified logic, for example, based on fuzzy production rules [2]. The possibilities of using intelligent systems with fuzzy rules for big data processing are described in a number of publications [3-4].

It should be noted that production-type logic is used not only in traditional for it expert systems [5], but also in a number of other areas of computer science. This circumstance is explained by that in fact a lot of models in computer science are of production character [6], and the structures for presenting data and knowledge are often hierarchical [7]. In particular, with the help of productions it is possible to describe

functional dependencies in relational databases calculated using inference based on Armstrong's axioms [8]. These dependencies play an important role in designing the structure of databases, especially «big» ones. In the article [9], it was shown for the first time that the relations of generalization and type aggregation in object-oriented programming possess the properties of production-logical inference. In [10], similar properties were noticed and used to optimize term rewriting systems.

An effective formalism for constructing and studying data and knowledge models in various subject areas is provided by algebraic methods [9-11]. In particular, the formal methodology for managing knowledge and logical inference in production type knowledge systems is developed by the lattice-based algebraic theory of LP-structures (lattice production structures) [12-13]. It is intended for the formal solution of a number of important tasks related to production systems. These include equivalent transformations, verification, equivalent minimization of knowledge bases. Another achievement of the theory is the relevant backward inference method (LP-inference), which significantly reduces the number of calls to external information sources [14]. This result is especially relevant when processing big data and managing large volumes of knowledge. LP-inference is based on the apparatus of production-logical equations.

In this report an extended class of such equations for an algebraic model, the expressive capabilities of which cover logical inference systems with a fuzzy knowledge base, is introduced for the first time. Some of these equations properties useful for finding their solutions are proved. Thus, the advantages of the theory of LP-structures extend to a fuzzy logical inference.

2 Logical-production equations in the FLP-structure

In this section a new class of equations related to fuzzy LP-structures is introduced. The question of a method for solving these equations is considered. The starting points of the theory of lattices, fuzzy sets, binary relations and LP-structures are described in [1, 15-18].

Let a fuzzy relation R on an atom-generated lattice $\mathbb{F}$ be given. Let also $\mu_R(A,B) > 0$ take place for some elements $A,B \in \mathbb{F}$. Then B can be called an image of A, and A – a preimage of B in relation R. In the FLP-structure, each element of the lattice can have many images and preimages, and with a different degree of membership (value $\mu_R(A,B)$).

For a given $B \in \mathbb{F}$, the minimal preimage with relation R is such an element $A \in \mathbb{F}$ that $\mu_R(A,B) > 0$ and A are minimal in the sense that it does not contain any other $A_1 \in \mathbb{F}$ for which $\mu_R(A_1,B) > 0$.

Definition 1. An atom $x \in \mathbb{F}$ is called initial in a fuzzy relation R if there is not any pair $A,B \in \mathbb{F}$ such that $\mu_R(A,B) > 0$, x is contained in B and is not contained in A. An element X is called initial if all its atoms are initial. A subset $\mathbb{F}_0(R)$

(sometimes denoted $\mathbb{F}_0$) consisting of all the initial elements $\mathbb{F}$ is called the initial set of lattice $\mathbb{F}$ in relation R. The initial set $\mathbb{F}_0$ defined above forms a sublattice in $\mathbb{F}$.

Let $\bar{R}$ be a logical closure of a relation R [18]. Taking into consideration its structure [16], it is easy to verify that the sets $\mathbb{F}_0(R)$ and $\mathbb{F}_0(\bar{R})$ coincide.

Consider the equation

$$\bar{R}(X) = B, \tag{1}$$

where $B \in \mathbb{F}$ is the given element, $X \in \mathbb{F}$ is the unknown.

Definition 2. *An approximate solution* to equation (1) is any preimage of an element B in $\mathbb{F}_0$ (in relation $\bar{R}$). *The solution (exact)* (1) is any minimal preimage of element B in $\mathbb{F}_0$. *A general solution* to an equation is the set of all its solutions.

Equations of the form (1) will be called *production-logical equations* in a fuzzy LP-structure.

Remark 1. By definition, the exact solution to equation (1) is also approximate. Besides, an approximate solution always contains at least one exact solution.

The main circumstance that creates difficulties for the solution process (1) is that usually only a relation R is given. In the modelled intelligent system, it corresponds to a given set of productions – the knowledge base. But the solution is required to be found as a preimage of the right-hand side of (1) in relation $\bar{R}$ – logical closure of R. At that, the complete construction of a logical closure is impractical, since in practice an unacceptable amount of resources will be required, both computational and in the sense of occupied memory.

An additional factor complicating not only the solution methods (1), but also the very formulation of this problem is the fuzziness of relation R. Apart from minimality of the desired preimage X required in Definition 2, it is necessary to take into account its second characteristic, namely, the value of the membership function $\mu_{\bar{R}}(X, B)$.

Next, we clear out the question of how the general solution of equations of the form (1) changes by their right-hand sides union. More precisely, if it is possible for a fuzzy relation R to solve several equations with simpler right-hand sides instead of the original equation.

Theorem 1. Let $\{X_p\}$, $p \in P$ be a general solution to an equation of form (1) with the right-hand side B_1, and $\{Y_q\}$, $q \in Q$ – a general solution to an equation of the same type with the right-hand side B_2. Then the general solution to equation (2) is the set of all the elements of form $X_p \bigcup Y_q$ from which elements containing other elements of the same set are excluded.

The significance of Theorem 1 is that it opens up the possibility of solving equation (1) by simplifying, that is, replacing it with several equations, each of which contains a lattice atom on the right side. At that, in the left part of the equations, according to Section 2, instead of the original relation R, an equivalent canonical relation can be considered. These transformations constitute the basic method for solving equation (1).

3 Conclusion

In the present work, the apparatus of production-logical equations in a generalized LP-structure is defined and investigated, which expands the scope of application of this algebraic theory to fuzzy intelligent systems of production type.

The reported study was supported by RFBR project 19-07-00037.

References

1. Piegat, A.: Fuzzy Modeling and Control. Springer-Verlag, Berlin, Heidelberg (2001).
2. Ozdemir, Y., Alcan, P., Basligil, H., Dokuz, C.: Just-In Time Production System Using Fuzzy Logic Approach and a Simulation Application. Advanced Materials Research 445, 1029–1034 (2012). doi:10.4028/www.scientific.net/amr.445.1029.
3. Fernandez, A., Carmona, C.J., Del Jesus, M.J., Herrera, F.: A View on Fuzzy Systems for Big Data: Progress and opportunities. International Journal of Computational Intelligence Systems 9(1), 69–80 (2016).
4. Elkano, M., Galar, M., Sanz, J., Bustince, H.: CHIBD: A Fuzzy Rule-Based Classification System for Big Data Classification Problems. Fuzzy Sets and Systems 348, 75–101 (2018).
5. Buchanan, B., Shortliffe, E: Rule-Based Expert Systems: The MYCIN Experiments of the Stanford Heuristic Programming Project. Addison-Wesley (1984).
6. Maciol, A.: An application of rule-based tool in attributive logic for business rules modeling. Expert Systems with Applications 34(3), 1825–1836 (2008).
7. Oles, F.: An Application of Lattice Theory to Knowledge Representation. Theor. Comput. Sci. 249(1) 163–196 (2000).
8. Armstrong, W.: Dependency structures of database relationships. In: Proc. IFIP Congress 1974, pp. 580–583. North-Holland, Amsterdam (1974).
9. Makhortov, S.: LP Structures on Type Lattices and Some Refactoring Problems. Programming and Comput. Software 35, 183–189 (2009). doi: 10.1134/S0361768809040021.
10. Makhortov, S.: An Algebraic Approach to the Study and Optimization of the Set of Rules of a Conditional Rewrite System. Journal of Physics: Conference Series 973(1), 12066.1–12066.8 (2018). doi: 10.1088/1742-6596/973/1/012066.
11. Hájek, P., Valdes, J.: A generalized algebraic approach to uncertainty processing in rule-based expert systems (dempsteroids). Comput. Artif. Intell. 10(1), 29–42 (1991).
12. Makhortov, S.: Production-logic Relations on Complete Lattices. Automation and Remote Control 73, 1937–1943 (2012). doi: 10.1134/S0005117912110161.
13. Makhortov, S., Bolotova, S.: An algebraic model of the production type distributed intelligent system. Journal of Physics: Conference Series 1203(1), 12045.1–12045.8 (2019). doi: 10.1088/1742-6596/1203/1/012045.
14. Bolotova, S., Trofimenko, E., Leschinskaya, M.: Implementing and analyzing the multithreaded LP-inference. Journal of Physics: Conference Series 973(1), 12065.1–12065.12 (2018). doi: 10.1088/1742-6596/973/1/012065.
15. Birkhoff, G.: Lattice Theory. Am. Math. Soc., Providence (1984).
16. Garmendia, L., Del Campo, R., López, V., Recasens, J.: An Algorithm to Compute the Transitive Closure, a Transitive Approximation and a Transitive Opening of a Fuzzy Proximity. Mathware & Soft Computing 16, 175–191 (2009).
17. Hashimoto, H.: Reduction of a Nilpotent Fuzzy Matrix. Information Sciences 27, 233–243 (1982).
18. Makhortov, S.: An Algebraic Model of the Intelligent System with Fuzzy Rules. Programmnaya Ingeneria 10(11–12), 457–463 (2019). doi: 10.17587/prin.10.457-463 (In Russian).

A Cloud-Native Serverless Approach for Implementation of Batch Extract-Load Processes in Data Lakes (Extended Abstract)

Anton Bryzgalov[1[0000-0001-9969-2251]] and Sergey Stupnikov[2[0000-0003-4720-8215]]

[1] Lomonosov Moscow State University, Moscow, Russia
tony.bryzgaloff@gmail.com

[2] Institute of Informatics Problems, Federal Research Center "Computer Science and Control" of the Russian Academy of Sciences, Moscow, Russia
sstupnikov@ipiran.ru

Data integration stands for the process of creating a unified view on the basis of multiple data sources. One can distinguish two ways for retrieving the required data, these are *virtual* and *materialized* integration approaches [1]. The virtual integration presumes the data are stored in the local sources and retrieved on demand. In such a way a user always get an up-to-date answer to the query. Materialized data integration paradigm requires the creation of a physical integrated data repository. One of the most widely used materialized integration techniques is *data warehousing* [2].

To manage data on their way from sources to the end user, data warehouses have to be accompanied with so-called Extract, Transform, and Load (ETL) periodical processes. A data warehouse is often built over a *data lake*. A data lake approach keeps the data minimally transformed comparing to the data sources. A data lake is usually populated using EL (Extract and Load) approach and the transformation is executed right after the data is migrated from the data lake to a data warehouse (ELT).

Several ELT-related approaches are known. ELTA approach in [5] stands for Extract, Load, Transform and Analyze where Analyze phase "makes business users efficiently utilize preprocessed data to under-stand enterprise behavior through analysis". Another ELT approach is proposed in [6] and is called *E-Hub*. E-Hubs are neutral Internet-based intermediaries that focus on specific industry verticals or specific business processes which provide secure trading environments to link with external buyers and suppliers. The implementation of proposed E-Hub architecture is based on Oracle Data Integrator. Authors of [7] present an ETL approach for near real-time data ingestion with focus on every step including Extraction. The approach challenges the way data changes are captured for frequent update and loaded without becoming overloaded and disrupting normal operational activities.

Both ETL and ELT processes can be represented in a form of a *workflow*, so *Workflow Management Systems* (WfMSs) are often considered as a part of the enterprise data infrastructure. Another important concern for modern enterprise solutions is the usage of public clouds which can dramatically reduce the infrastructure and management costs. However, to use the cloud resources efficiently an application has to meet the cloud-native design principles. Maximum utilization of cloud resources can be

achieved by applying the *serverless paradigm* [4] which means that the resources are allocated only at the time of actual code execution and teared down when the execution is finished. So serverless WfMSs are being developed [8].

Nevertheless, Big data analytics systems review [3] shows that current ELT and ETL solutions are often based on massive parallel processing architectures and are tightly connected with the hardware they are based on. Purely serverless EL solutions are not known.

This work advocates new approach to deal with batch EL processes for data lakes. From a high-level perspective, the approach includes three main steps: data extraction tasks execution, data partitioning and deduplication. As a result, the data is stored in a persistent storage and can be further used by outer data processing infrastructure. The proposed approach is aimed to fit use cases which are characterized by (1) discontinuous batch jobs (2) that are executed in a cloud infrastructure. EL processes are run according to a schedule and execution time of a single run is less than an interval between the sequential executions. This causes existence of idle periods for infrastructure which leads to underutilization of resources and leaves an opportunity to apply serverless approach. Utilization of resources is increased by shutting them down after an execution has finished.

A configurable architecture dealing with batch EL processes in a cloud infrastructure is proposed (**Fig. 1**.). The system extracts the data from the external data sources and stores it to a data lake. The system operates with tasks which are atomic execution units and describe single portions of data to be managed. The *task lifecycle* begins from its creation by *tasks creator* component. Tasks creator reads the configuration file provided by the user, fetches credentials from *credentials storage* and sends the requests to the source system to populate the *tasks queue* with the tasks description. A task «flows» through all the executing components as it is shown via white thick arrows. After the tasks queue is populated the tasks creator component is shut down and the *tasks executor* component (yellow badge) starts to execute tasks by sending requests to the source systems and storing the received data to the storage. The tasks executor stores the downloaded data to the *data landing zone*. The tasks creator and tasks executor have to send requests to the source system with respect of the source system rate limits. These can be achieved by throttling the requests using the local *throttlers* which are components aimed to avoid violating the sources rate limits. When all the tasks are executed then the tasks executor component stops its work and the tasks data partitioner and deduplicationer starts to merge the data from the landing zone to the persistent data location. The partitioning and deduplication are based on tasks metadata which is saved in the tasks queue. The tasks creator and tasks executor components are serverless and therefore have to store all the execution status-related data in an external storage to provide fault tolerance. On the basis of DAGMan Design Principles [9] a task lifecycle algorithm is defined accompanied with per-task finite state machine (FSM). Overall time complexity of the loop-based algorithm is $O(N)$, where N is the number of initially pending tasks.

The described architecture is implemented in a form of a serverless data application prototype based on Amazon Web Services (AWS).

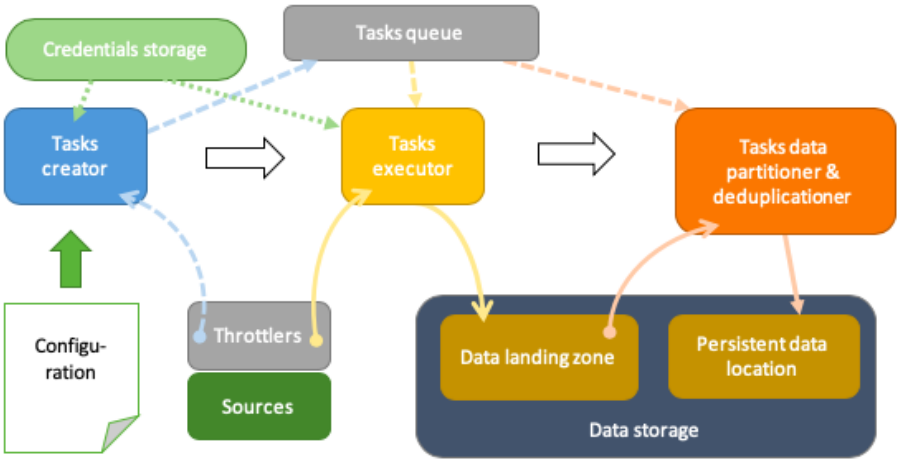


Fig. 1. Extract-Load System Architectural Scheme

Table 1. Cloud services representing architectural components

Component	Cloud service
Tasks creator	AWS Lambda
Tasks executor	AWS Batch
Tasks partitioner & deduplicationer	AWS Batch
Tasks queue	AWS Dynamo DB
Throttlers	AWS Dynamo DB
Credentials storage	AWS Secrets Manager
Data Storage	AWS S3

The correspondences between the cloud services and the architecture components are presented in **Table 1**. The execution is triggered by AWS CloudWatch. The service triggers execution of the state machine inside AWS Step Functions which handles the order of the tasks creator, tasks executor and tasks partitioner & deduplicationer components execution.

The prototype has been evaluated against an enterprise EL process involving Google Analytics as a source system. Using the proposed approach reduces the infrastructure costs by 86% and increases the resources utilization by 74%.

The described architecture is not limited to Amazon Web Services as the only supported cloud provider. The used services have analogues in other clouds like Azure and Google Cloud Platform.

One of the main discussion points of the approach is the serverless paradigm limitations. First risk is vendor lock: all the serverless resources are fully managed and cannot be easily migrated to another cloud provider unless some cross-cloud framework is used. Another limitation is unpredictable resources availability: shared execu-

tion environment may need to wait up to several minutes to find available execution slots to execute the requested job. When the resources are payed on-demand or even reserved then they are available all the time. But this requires resources maintenance and may cause underutilization. A negative effect of serverless paradigm is a cold start effect. This is a typical behavior of FaaS services: the first run of a function requires time for environment startup. But in terms of long-running data application jobs this is a minor drawback.

Acknowledgments

The research is financially supported by the Russian Foundation for Basic Research, projects 18-07-01434, 18-29-22096.

References

1. Lenzerini, M.: Data Integration: A Theoretical Perspective. In: Proceedings of the twenty-first ACM SIGMOD-SIGACT-SIGART symposium on Principles of database systems, Madison, Wisconsin, 2002, pp. 233–246. https://dl.acm.org/doi/10.1145/543613.543644
2. Calvanese, D., de Giacomo, G., Lenzerini, M., Nardi, D.: Data Integration in Data Warehousing. In: International Journal of Cooperative Information Systems, Vol. 10, No. 3, 2001, pp. 237–271.
3. Elgendy, N., Elragal, A.: Big Data Analytics: A Literature Review Paper. In: Perner P. (eds) Advances in Data Mining. Applications and Theoretical Aspects. ICDM 2014. Lecture Notes in Computer Science, vol 8557. Springer, Cham. https://doi.org/10.1007/978-3-319-08976-8_16
4. Kim, Y., Lin, J.: Serverless Data Analytics with Flint. 2018 IEEE 11th International Conference on Cloud Computing (CLOUD), San Francisco, CA, 2018, pp. 451–455. https://doi.org/10.1109/CLOUD.2018.00063
5. Marín-Ortega, P., Dmitriyev, V., Abilov, M., Gómez, J.: ELTA: New Approach in Designing Business Intelligence Solutions in Era of Big Data. In: Procedia Technology, 16, 2014. https://doi.org/10.1016/j.protcy.2014.10.015
6. Zhou, G., Xie, Q., Hu, Y.: E-LT Integration to Heterogeneous Data Information for SMEs Networking Based on E-HUB. In: Fourth International Conference on Natural Computation, Jinan, 2008, pp. 212-216, https://doi.org/10.1109/ICNC.2008.77
7. Sabtu, A., et al.: The challenges of Extract, Transform and Loading (ETL) system implementation for near real-time environment. In: 2017 International Conference on Research and Innovation in Information Systems (ICRIIS), Langkawi, 2017, pp. 1-5, https://doi.org/10.1109/ICRIIS.2017.8002467
8. Jiang Q., Lee Y.C., Zomaya A.Y.: Serverless Execution of Scientific Workflows. In: Maximilien M., Vallecillo A., Wang J., Oriol M. (eds) Service-Oriented Computing. ICSOC 2017. Lecture Notes in Computer Science, vol 10601, 2017. Springer, Cham. https://doi.org/10.1007/978-3-319-69035-3_51
9. Couvares P., Kosar T., Roy A., Weber J., Wenger K.: Workflow Management in Condor. In: Taylor I.J., Deelman E., Gannon D.B., Shields M. (eds) Workflows for e-Science, 2007. Springer, London. https://doi.org/10.1007/978-1-84628-757-2_22

Denotative and Significative Semantics Analysis and Control Methods in Diagrams (Extended Abstract)

Nikolay Voit[1][0000-0002-4363-4420] and Semen Bochkov[0000-0003-1089-4119]

[12] Ulyanovsk State Technical University, Severny Venets str. 32, 432027 Ulyanovsk, Russia
n.voit@ustu.ru, bochkovsi@ido.ulstu.ru

Abstract. Authors proposed algorithms for the analysis and control of the design workflows diagrammatic models' denotative and significative semantics. The denotative characteristics are the basis for the denotation description as a graphic language concept; therefore, a dictionary of such graphic words is formed in the work. These characteristics allow to determine the synonyms and antonyms of graphic words, due to which it is possible to eliminate errors in understanding the diagram operation (semantic errors). Significative semantics analysis reveals structural errors based on isomorphism and homomorphism of workflow traces.

Keywords: Workflows, Graphic Languages, Antonymy, Synonymy.

1 Introduction

During CAD systems design, diagrammatic models are actively used, presented in artifacts of visual graphic languages BPMN, UML, IDEF and others. This significantly increases the design process efficiency and the quality of the created systems due to the process participants' interaction language unification, rigorous documentation of design and architectural, functional solutions and formal control of diagram correctness.

At the same time, an obligatory step in enterprise business processes modeling is automatic and/or automated verification of the obtained models. Defect-free completion analysis issues are relevant and actual, since the complexity of models is constantly increasing, and the verification tools built into the simulation environment are far from perfect.

The variety of diagrammatic graphic languages covers all possible system descriptions types; however, unsolved problems exist. Graphic design support tools do not use universal methods of parsing and are highly specialized and aimed at working with few graphic languages. The control tools development for new languages takes considerable time, because it actually requires a new analyzer creation from the scratch. Known methods of parsing have significant costs in time (exponential, polynomial characteristics) and memory.

2 Related Works

The Common Workflow Language (CWL) workflows descriptors are researched in [1], which allow analyzing data in various computing environments (Docker containers virtualization). The authors developed CWL-metrics, a utility tool for cwltool (a reference implementation of CWL), for collecting Docker container runtime metrics and workflow metadata for analyzing resource requirements. To demonstrate the use of this tool, the authors analyzed 7 workflows on 6 types of instances. Analysis results show that choosing the type of instance allows to reduce financial costs and execution time using the required amount of computing resources. However, in the CWL-metrics implementation proposed by the authors, there are no functions for collecting metric data from parallel tasks in workflows.

[2] is devoted to the dynamic access control approach for business processes development. Authors offer a context-oriented and trust-oriented work environment. The proposed approach focuses on inter-component relationships, where steps are performed online or offline to avoid performance bottlenecks. It should be noted that the presented context-oriented access structure is applicable only for solving problems related to business processes in service-oriented computing.

[3] described a new approach to the systematic support of engineers using model-driven system architectures for process design and plant automation. The authors investigated a new aspect the virtual intelligent objects design in enterprise data models, which represents the life cycle of an object. A methodology is described that enables users to define the life cycle for classes of objects depending on the context and goals of the projects. The authors performed workflows research and analysis to form a library of production processes for certain objects classes. However, the dynamic distributed workflows analysis and control methods are not considered.

In [4], an approach to the selection of services for modeling business processes is proposed. At the first stage, the function similarity method is used to select services from the service repository to create a set of candidate services that checks the description of functions to find suitable services, especially a service can publish one or more functions through several interfaces. At the second stage, a method based on the probabilistic model verification, which includes the composition of services and calculation of stochastic behavior in accordance with the workflow structures, is used to quantitatively verify process instances. Next, experiments are carried out to demonstrate the efficiency and effectiveness of the proposed method compared to traditional methods.

In [5], an improved two-stage approach of exact query based on the graph structure is proposed. At the filtering stage, a composite task index, which consists of a label, a connection attribute and a task attribute, is used to obtain candidate models, which can significantly reduce the number of process models that need to be tested at a specific time - the verification algorithm. At the verification stage, a new subgraph isomorphism test based on the task code is proposed to clarify the set of candidate models. The experiments are conducted on six synthetic model kits and two real model kits. However, the algorithm has polynomial computational complexity.

In [6], success and failure factors for the implementation of business process management technologies in organizations were investigated.

3 Mathematical Apparatus

In CAD systems, hybrid dynamic workflows diagrammatic models denotative semantics control and analysis methods cooperate with synonyms and antonyms of temporal words in graphical languages. The goal is to detect errors in diagrammatic models events and thereafter correct them.

The method differs from similar ones in functioning with hybrid dynamic diagrammatic design workflows models, temporal graphic words of denotative semantics. The method also has a linear law of analysis time complexity. The hybrid dynamic design workflows diagrammatic models control and analysis method in denotative semantics contains procedures for diagrammatic models and denotative semantics analysis.

Consider the procedure for analyzing a hybrid dynamic diagrammatic model (diagram) of design workflows. The input is a diagram; its model has the following view: G = (V,E,TV,TE), where:

- V is vertices set,
- E is edges set, $E \subset (V \times V)$,
- TV is vertices types set,
- TE is edges types set.

The output is correctness status of diagram with an error message if exists.

Hybrid dynamic workflows diagrammatic models significative semantics control and analysis methods works with isomorphism, homomorphism of temporal traces (tracks) of diagrammatic models of a graphic language in order to identify structural errors in diagrammatic models for subsequent transformations of these traces. The method differs from its analogues in that it works with hybrid dynamic design workflows diagrammatic models, temporal graphic words of significative semantics and has a linear law of time complexity of analysis.

Hybrid dynamic workflows diagrammatic models significative semantics control and analysis methods contain procedures for analyzing diagrammatic models and significative semantics.

For the traversing process management define a finite state machine A of RVTI-grammar [7] in the following form: A = (S, T, S_0, C, Send, Ftrans), where

- S is states set,
- T is input chars (terms) set,
- S_0 is an initial state,
- C is transitions conditions set,
- FA is transitions functions set,
- Send is finite states set,
- Ftrans is a transition function: S x T x C -> S x FA

4 Conclusion

The proposed RVTI-grammar and methods, software and information tools are the contribution and development of the following science areas:

1. Theory and practice of design, development and maintenance of automated systems.
2. The formal languages and grammars theory.
3. Theory and practice of processing visual and graphic languages.

New analysis and control methods in denotative and significant semantics and semantics of diagrammatic design workflows models have been developed. They provide the quality improvement of such diagrams and expand the theoretical foundations of the business process management theory.

Scientifically, the author's RVTI-grammar and analysis methods provide a linear time law for the analysis of the diagram because of the automaton approach and analytical proof of its usage.

In practice, the author's RVTI-grammar and methods improve the diagrams quality identifying complex semantic errors at the conceptual stage of technical development, which provides an economic positive effect.

The outlook is interpretation methods of these diagrams in various bases of graphic languages, including the UML, BPMN etc.

References

1 Ohta, T., Tanjo, T., Ogasawara, O.: Accumulating computational resource usage of genomic data analysis workflow to optimize cloud computing instance selection. GigaScience, 8 (4), 1-11 (2019).

2 Bhattasali, T., Chaki, N., Chaki, R., Saeed, K.: Context and trust aware workflow-oriented access framework. In: Proceedings of the International Conference on Software Engineering and Knowledge Engineering, SEKE-2016, http://ksiresearchorg.ipage.com/seke/seke16paper/seke16paper_179.pdf, last accessed 2020/06/01.

3 Bigvand, G., Fay, A.: A workflow support system for the process and automation engineering of production plants. In: 2017 IEEE International Conference on Industrial Technology (ICIT), pp. 1118-1123 (2017).

4 Gao, H., Chu, D., Duan, Y., Yin, Y.: Probabilistic Model Checking-Based Service Selection Method for Business Process Modeling. International Journal of Software Engineering and Knowledge Engineering, 6 (27), 897–923 (2017).

5 Huang, H., Peng, R., Feng, Z.: Efficient and Exact Query of Large Process Model Repositories in Cloud Workflow Systems. IEEE Transactions on Services Computing, 11 (5), 821-832 (2018).

6 Reijers, H.A., Vanderfeesten, I., van der Aalst, W.M.P.: The effectiveness of workflow management systems: A longitudinal study. International Journal of Information Management, 36 (1), 126-141 (2016).

7 Afanasyev, A.N., Voit, N.N. & Kirillov, S.Y.: Temporal Automata RVTI-Grammar for Processing Diagrams in Visual Languages as BPMN, eEPC and Askon-Volga. Proceedings of the 2019 5th International Conference on Computer and Technology Applications - ICCTA 2019. Available at: http://dx.doi.org/10.1145/3323933.3324067 (2019).

An Ontology Approach to Data Integration (Extended Abstract)

Manuk Manukyan

Yerevan State University, Yerevan 0025, Armenia
mgm@ysu.am

Abstract. In the frame of an XML-oriented data model an ontology approach to data integration is considered. Three kinds mechanisms are used to formalize data integration concept: reasoning rules, content dictionaries and semantical constraints. We introduced the concept an integrable data as ontology object which has been formalized within so-called algebra of integrable data. The offered ontology is presented by reasoning rules which are based on the proposed mathematical model. Mappings from source data models into ontology are defined by means of algebraic programs. To support the conceptual entities an algorithm to generate mappings from relational data sources into ontology is developed.

Keywords: Data Integration, Data Warehouse, Mediator, Data Cube, Ontology, Reasoning Rules, Metamodel, XML, OPENMath.

1 Introduction

In this paper we will consider an approach to ontology-based data integration. Ontology in informatics is understood as a formal knowledge representation in the form of a set of concepts of some subject domain and relations between them. Such representations are used for reasoning about entities of the subject domains, as well as for the domains description [2]. Various metamodels have been developed to define ontology [2, 5]. We consider an XML-oriented data model which is a result of strengthening the XML data model by means of the OPENMath concept [1] as a metametamodel to support an ontology approach to data integration. OPENMath is a formalism to represent mathematical concepts with their semantics and is implemented as an XML application. We have certain experience in OPENMath usage in our research to support databases with ontological dependencies [3].

2 A Formalism for Defining Metamodels

We consider the XML-oriented data model as a best solution to use as metamodels construction formalism after addressing some of the shortcomings. Choosing the XML data model as a metamodel construction formalism is explained with the fact

that this model is some compromise between conventional and semi-structured DMs because in contrast to:

– semi-structured DM, the concept of database schema in the sense of conventional DMs is supported;

– conventional DMs hard schemas, there is possibility to define more flexible database schemas.

The weakness of XML data model is the absence of data types concept in conventional sense. To eliminate this shortcoming and to support ontological dependencies on the XML data model level, we expand the XML data model by means of the OPENMath concept. The result of such extension is a data model which coincides with XML data model and which was strengthened with computational and ontological constructs of OPENMath.

3 Algebra of Integrable Data

In the frame of our approach to ontology-based data integration we are introducing the concept of integrable data as an entity of the conceptual level [4]. The interpretation of the integrable data concept in the frame of the metamodel has been proposed.

3.1 Operations

Virtual and materialized integration of data assumes introduction of special operations, such as filtering, joining, aggregating, etc. The proposed operations are analogs of the corresponding operations in realtional algebra.

4 Data Integration ontology

Our approach to ontology-based data integration concept assumes formalizing the mediator, data warehouse and data cube concepts by an XML DTD. Formalization result of these concepts are so-called ontology reasoning rules. These rules will be interpreted by means of algebraic programs.

4.1 Mathematical Model

In our case research object is the area of ontology-based data integration. We should formalize entities of this subject domain and define relationships between these mathematical objects. The proposed formalization will be the mathematical basis for constructing the reasoning rules. We differentiate three kind of reasoning rules: mediator rule, data warehouse rule and data cube rule. A mathematical model of the suggested concept of ontology-based data integration is constructed.

4.2 Ontology as an XML Application

Based on the above discussed formalisms, an XML application to support conceptual schemas during ontology-based data integration is developed. Namely, an XML DTD

was constructed based on the proposed mathematical relationships. In Appendix A an XML DTD for modeling the reasoning rules is presented. The proposed XML DTD is an instance of the above considered metamodel which is an advanced XML data model and we use it as an ontology definition language. In contrast to OWL which is a description logic based ontology language for semantic Web, the considered metamodel is oriented to data integration and is based on an algebra of integrable data.

5 Mappings Generation

Our concept to ontology-based data integration assumes constructing a mapping from arbitrary source data model into ontology. In the frame of this paper we consider some issues when constructing the mapping from relational source data into ontology. An algorithm to generate a query to extract data from relational data sources is proposed. The proposed algorithm is based on the in-order method of the tree traversal. We are using the proposed algorithm to support the data warehouse and data cube concepts. In the mediator case, this algorithm should be modified. The detailed discussion of problems to support queries over the mediator is beyond the topic of this paper.

6 Conclusions

In the frame of an approach to ontology-based data integration, an ontology definition metamodel is proposed. The considered metamodel is oriented to XML data model which has been extended by the OPENMath concept. In the result of such extension, the XML data model has been strengthened with ontological and computational constructions. We introduced the concept of an integrable data as an ontology object which has been formalized within so-called integrable data algebra developed by us. The considered algebraic operations are analogs of the correponding operations in relational algebra. The interpretation of the integrable data concept in the frame of the metamodel has been proposed. A mathematical model of the suggested concept of ontology-based data integration is constructed. The concepts of the integrable data algebra have been formalized by mechanisms of content dictionary and signature files (semantical constraints) of the OPENMath. The proposed ontology for data integration concept is presented by reasoning rules which are based on the considered mathematical model. The offered XML DTD is an instance of the considered metamodel. Thus, the ontology-based data integration concept formalization result is an ontological modeling language which is defined as an XML application. Mappings from source data models into ontology are defined by means of algebraic programs. To support the conceptual entities, an algorithm to generate mappings from relational data sources into ontology is developed. The output of this algorithm is an SQL-program by means of which we can extract data from relational sources to support the concepts of data warehouse and data cube. It is essential, that the metamodel is extensible, which allows to integrate arbitrary data models by using a computationally complete language.

Acknowledgments

This work was supported by the RA MES State Committee of Science, in the frames of the research project No. 18T-1B341.

References

1 Drawar, M. OpenMath: An Overview. *ACM SIGSAM Bulletin*, 34(2):2–5, 2000.
2 Kalinichenko, L. A. Effective Support of Databases with Ontological Dependencies: Relational Languages instead of Description Logics, *Programmirovanie*, 38(6):315–326, 2012.
3 Manukyan, M. G. On an Ontological Modeling Language by a Non-Formal Example. In *CEUR-WS*, volume 2277, pages 41–48, 2018.
4 Manukyan, M. G. Ontology-based Data Integration. In *CEUR-WS*, volume 2523, pages 117–128, 2019.
5 OMG. Ontology Definition Metamodel. In *OMG Specification*, 2014.

An XML DTD for Modeling the Reasoning Rules

```
<!– – include dtd for extended OPENManth objects – –>
<!ELEMENT dir (source+, (med | whse | cube)) >
<!ELEMENT med (msch)+ >
<!ELEMENT msch (sch, wrapper) >
<!ELEMENT sch (OMATTR) >
<!ELEMENT wrapper (OMA) >
<!ELEMENT whse (wsch, extractor) >
<!ELEMENT wsch (OMATTR) >
<!ELEMENT extractor (OMA) >
<!ELEMENT cube (ssch, mview) >
<!ELEMENT ssch (OMATTR) +>
<!ELEMENT mview (view+, granularity+) >
<!ELEMENT view (OMA) >
<!ELEMENT granularity (partition)+>
<!ELEMENT partition EMPTY >
<!ELEMENT source (OMATTR)+ >
<!ATTLIST source name CDATA #REQUIRED >
<!ATTLIST granularity name CDATA #REQUIRED >
<!ATTLIST partition name CDATA #REQUIRED >
<!ATTLIST view name CDATA #REQUIRED >
```

DATA MANAGEMENT IN SEMANTIC WEB

Pragmatic interoperability and translation of industrial engineering problems into modelling and simulation solutions (Extended Abstract)

Martin T. Horsch[1], Silvia Chiacchiera[1], Michael A. Seaton[1], Ilian T. Todorov[1], Björn Schembera[2], Peter Klein[3], and Natalia A. Konchakova[4]

[1] UK Research and Innovation, STFC Daresbury Laboratory, Keckwick Ln, Daresbury, Cheshire WA4 4AD, UK
{martin.horsch, silvia.chiacchiera, michael.seaton, ilian.todorov}@stfc.ac.uk
[2] Höchstleistungsrechenzentrum Stuttgart, Nobelstr. 19, 70569 Stuttgart, Germany
schembera@hlrs.de
[3] Fraunhofer-Institut für Techno- und Wirtschaftsmathematik, Fraunhofer-Platz 1, 67663 Kaiserslautern, Germany
peter.klein@itwm.fraunhofer.de
[4] Helmholtz-Zentrum Geesthacht, Magnesium Innovation Centre, Max-Planck-Str. 1, 21502 Geesthacht, Germany
natalia.konchakova@hzg.de

The capabilities of service, software, and data architectures increase greatly if they are able to integrate a variety of heterogeneous resources into a common framework. Interoperability, which permits integrating distributed and heterogeneous infrastructures, requires an agreement between platforms and components with respect to aspects pertaining to three major areas of the theory of formal languages: Syntax, semantics, and pragmatics – or how to write correctly (according to a given format or grammar), how to associate a meaning with the communicated content (by which data items become information), and *how to deal with information* and transactions that involve an exchange of information. While well-known and well-established approaches for ensuring syntactic and semantic interoperability exist, *pragmatic interoperability* has not acquired the same degree of attention.

Software and data architectures often neglect to explicitly formulate any requirements at the level of pragmatics, since they are assumed to be guaranteed by institutional procedures (e.g., who is given an account, and who may ingest data). However, this delegation of responsibilities cannot be upheld for *open infrastructures* where anybody is invited to participate and to which a multitude of external tools and platforms connect, each of which may have its own users, roles, service definitions, access regulations, interfaces, and protocols. Accordingly, finding that semantic interoperability cannot reach its goals if it is not supplemented by an agreement on "what kind of socio-technical infrastructure is required," it has been proposed to work toward a universal *pragmatic web* [6]; in full consequence, this would add a third layer to the world-wide web infrastructure, operating on top of the semantic web and hypertext/syntactic web layers. This raises the issue of requirements engineering (i.e., specifying and implement-

ing requirements) for service-oriented infrastructures, which becomes non-trivial whenever "stakeholders do not deliberately know what is needed" [9]. Previous work has established that ontologies are not only a viable tool for semantic interoperability, but also for enriching the structure provided for the semantic space by definitions of entities, relations, and rules that are employed to specify jointly agreed pragmatics [6, 8].

The present work follows a similar approach; it intends to contribute to the aim of the European Materials Modelling Council (EMMC) to make services, platforms, and tools for modelling and simulation of fluid and solid materials interoperable at all levels, which includes pragmatic interoperability. The workflow pattern standard of the EMMC is MODA (i.e., Model Data) [1], which as an ontology becomes OSMO, the ontology for simulation, modelling, and optimization; this ontology development [4] constitutes the point of departure for the present discussion. One of the concepts at the core of this line of work is that of materials modelling *translation*, i.e., the process of guiding an industrial challenge toward a solution with the help of modelling [2, 3]. The experts that facilitate this process are referred to as *translators*; they provide a service for companies and can be either academics, software owners, internal employees of a company, or independent engineers.

The present release of the MMTO (version 1.3.4) and OSMO (version 1.6.6) is available as free software under the terms and conditions of the GNU Lesser General Public License version 3 [5, 7]. The relevant part of the MMTO and OSMO class hierarchy is visualized in Fig. 1, including the relations that are most useful and common in this context. The translation case *aspects*, cf. Tab. 1, directly correspond to the text fields from the EMMC Translation Case Template [2], except that the MMTO permits the provision of semantically characterized content; this follows the approach from OSMO, which delivers the same feature for the text fields from MODA.

Acknowledgment

The authors thank N. Adamovic, W. L. Cavalcanti, G. Goldbeck, and A. Hashibon for fruitful discussions. The co-author P.K. acknowledges funding from the European Union's Horizon 2020 research and innovation programme under grant agreement no. 721027 (FORCE), the co-author N.K. under grant agreement 723867 (EMMC-CSA), and the co-authors S.C., M.T.H., M.A.S., and I.T.T. under grant agreement no. 760907 (Virtual Materials Marketplace). The present work was facilitated by activities of the Innovation Centre for Process Data Technology (Inprodat e.V.), Kaiserslautern.

References

1. CEN-CENELEC Management Centre: Materials modelling: Terminology, classification and metadata. CEN workshop agreement 17284, Brussels, Belgium (2018)
2. EMMC Coordination and Support Action: EMMC Translation Case Template. https://emmc.info/emmc-translation-case-template/ (2017), date of access: 31st December 2019

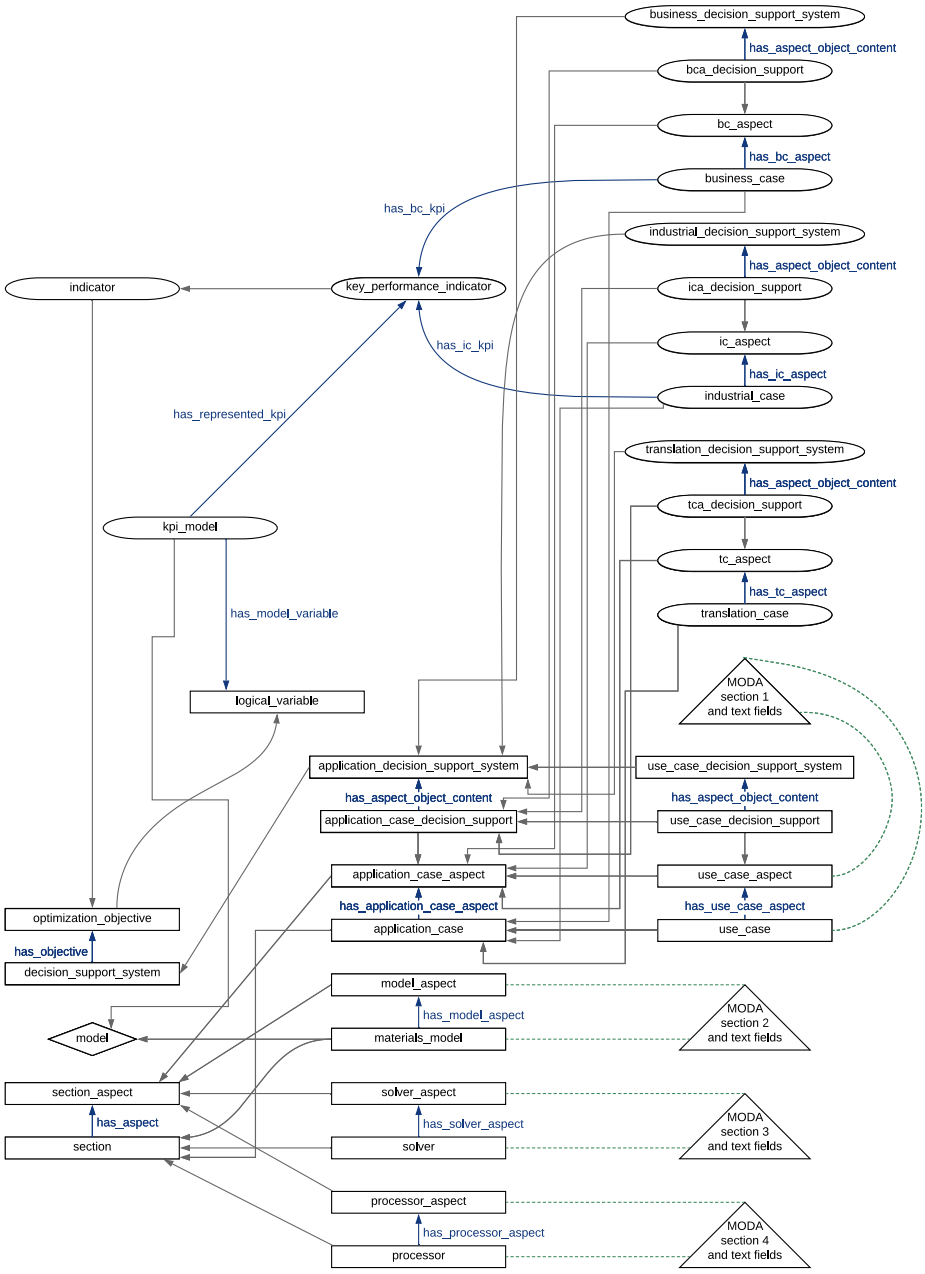


Fig. 1. Entities from MMTO version 1.3.4 (rounded boxes), OSMO version 1.6.6 (rectangular boxes), EMMO version 1.0.0 alpha 2 (diamond), MODA sections (triangles), and a subset of the relations defined by MMTO and OSMO, where grey arrows correspond to the transitive reduction of the `rdfs:subClassOf` relation, blue arrows to relations between individuals as indicated by arrow labels, and green dashed lines to correspondences between elements from MODA and OSMO.

Table 1. Aspects of a translation case (TC), `mmto:translation_case`, specified by the MMTO on the basis of the EMMC Translation Case Template [2].

aspect class name	content description
`mmto:tca_translator`	*translator(s) involved in the TC* content type: `evmpo:translator` [5]
`mmto:tca_end_user`	*involved end user(s), i.e., client(s) of the translator* content type: `evmpo:end_user` [5]
`mmto:tca_industrial_case`	*industrial case(s) associated with the TC* content type: `mmto:industrial_case`
`mmto:tca_business_case`	*business case(s) associated with the TC* content type: `mmto:business_case`
`mmto:tca_expected_outcome`	*expected outcome of the translation process* content type: plain text, i.e., `xs:string`
`mmto:tca_pe_type`	*physical equation type(s) employed for modelling* content type: `osmo:physical_equation_type` [4]
`mmto:tca_discussion`	*summary of discussions with the end user* content type: plain text, i.e., `xs:string`
`mmto:tca_kpi_model`	*employed key performance indicator model(s)* content type: `mmto:kpi_model`
`mmto:tca_evaluation`	*evaluation (assessment) of the TC* content type: `vivo:translation_assessment` [5]
`mmto:tca_impact`	*impact and benefit to the end user; how does the TC contribute to improving processes/products?* content type: plain text, i.e., `xs:string`
`mmto:tca_decision_support`	*employed decision support system(s)* content type: `osmo:decision_support_system`

3. EMMC Coordination and Support Action: EMMC Translators' Guide. https://emmc.info/translators-guide-2/ (2018), date of access: 31st December 2019
4. Horsch, M.T., Niethammer, C., Boccardo, G., Carbone, P., Chiacchiera, S., Chiricotto, M., Elliott, J.D., Lobaskin, V., P. Neumann, P.S., Seaton, M.A., Todorov, I.T., Vrabec, J., Cavalcanti, W.L.: Semantic interoperability and characterization of data provenance in computational molecular engineering. Fluid Phase Equilib. **65**(3), 1313–1329 (2020)
5. Horsch, M.T., Chiacchiera, S., Seaton, M.A., Todorov, I.T., Toti, D., Goldbeck, G.: Introduction to the VIMMP ontologies. Technical report, EMMC ASBL, Brussels, Belgium (2020). https://doi.org/10.5281/zenodo.3936796
6. Schoop, M., de Moor, A., Dietz, J.: The pragmatic web: A manifesto. Comm. ACM **49**(5), 75–76 (2006)
7. VIMMP Project Consortium: Virtual Materials Marketplace (2020), https://vimmp.eu/, date of access: 15th July 2020
8. Weidt Neiva, F., David, J.M.N., Braga, R., Campos, F.: Towards pragmatic interoperability to support collaboration: A systematic review and mapping of the literature. Informat. Software Technol. **72**, 137–150 (2016)
9. Wiesner, S., Thoben, K.D.: Requirements for models, methods and tools supporting servitisation of products in manufacturing service ecosystems. Int. J. Comput. Integr. Manuf. **30**(1), 191–201 (2016)

Analysis of the Semantic Distance of Words in the RuWordNet Thesaurus (Extended Abstract)

Liliya Usmanova [1[0000-0002-6784-4035]], Irina Erofeeva [1[0000-0002-6702-9129]], Valery Solovyev [1[0000-0003-4692-2564]], Vladimir Bochkarev [2[0000-0001-8792-1491]]

[1] Institute of Philology and Intercultural Communication, Kazan Federal University, Kazan, 18 Kremlyovskaya street, 420008, Russian Federation
[2] Institute of Physics, Kazan Federal University, Kazan, 18 Kremlyovskaya street, 420008, Russian Federation
usmanova77@rambler.ru

The issue of distinguishing between synonyms and non-synonyms remains undecided in linguistics, and is solved at the level of the conceptual layer of semantics, depending on the dominant scientific paradigm. Synonymy within the framework of system centrism is considered a hierarchical combination of words of the same part of speech basing on the identity or similarity of their lexical meanings with clear boundaries between synonymic series, which enable to differentiate one synonymic series from another. In line with the cognitive approach, synonyms are considered the field structures or fragments of a continuous and fundamentally incomplete synonymic and, taken more broadly, associative-verbal network.

It is customary in linguistics to distinguish absolute synonyms (doublets), lexical synonyms and quasi-synonyms in accordance with the semantic distance parameter imposed on words of this kind.

Quasi-synonyms are defined as indirect, approximate synonyms; their meanings can vary in conceptual content, speaker's attitude, collocations, etc. and change depending on the context. Alongside with quasi-synonyms, some researchers tend to identify the so-called analogues, their meaning substantially intersecting with the general meaning of a given series of synonyms, although not reaching the degree of closeness to it, which constitutes synonymy itself. The most common examples of analogues are cohyponyms, hyperonyms and hyponyms, as well as lexical units associated with a given lexical unit with a less defined semantic connection can be also included among the analogues.

The object of our study has become a comparative analysis of analogues from the New Explanatory Dictionary of Synonyms of the Russian Language (NEDS) [1] for the degree of their distance in the RuWordNet thesaurus [2].

The following traditional linguistic methods, such as the descriptive method, the onomasiology method and the derivation analysis method were involved in the process of work, as well as the modern methods of cognitive and corpus linguistics.

A computer program was created to automate the process, which listed shortest routes between specified pairs of vertices for each pair of analogous words along the

semantic relations in the thesaurus. The proposed methodology for computer generation of the shortest paths between words in the thesaurus with subsequent substantial analysis of the path structure is new and can be used for thesaurus verification.

Basing on the obtained data, recommendations were formulated for RuWordNet to reduce the distances between analogues, and general analysis principles were proposed, which can be useful for verifying RuWordNet.

Further work was carried out using the following methods. The conditions for the approaching of analogues in the structure of the thesaurus were identified applying *component and definitive analysis* for establishing the basic meanings of the analyzed words, for identifying differential semes, as well as the signs leading to the conceptual domain (DOMAIN).

The application of method of *lexical-semantic fields indicating* made it possible to identify the relationships between the elements of the field in a paradigmatic plan. The analysis procedure consisted in:

1. Consideration of the vocabulary definition of the target word as a detailed definition: identifier word + specifying words;

2. The field constituents identification was carried out further on the basis of their recurrence in the dictionaries of the same identifiers;

3. The analysis of the relationships between the elements of the field in a paradigmatic respect: synonymic, hyper-hyponymic relationships, cognates, part-of-speech synonymy, etc.

The cognitive interpretation of meanings was based on the techniques of *conceptual analysis of words*, which implies transition from the content of meanings to the content of concepts (considering etymologic data). Conclusions about the significance of certain cognitive traits in understanding the concept were drawn basing on the predominance either of the generic or the specific characteristics.

Observations manifest that the considered pairs of analogues in terms of semantic relations are accessible in the structure of the thesaurus in a small number of steps. Out of the 2590 pairs of analogues obtained from NEDS by the continuous sampling method, 2548 words in RuWordNet are located within the distance of 4 or even fewer steps by the semantic relations, 39 – at a distance of 5, and 3 pairs – at a distance of 6 steps. The fact that 98% of the analogues are located at a close distance in the thesaurus is a serious argument in favor of a good organization of the structure of the thesaurus. The 4 steps as a measure of proximity are chosen so to say provisionally to a certain degree, based on the agreed intuition of the project executors. A long distance (5 steps or more) may be an indication of errors in the thesaurus or a consequence of an imperfect analytic algorithm.

The basis of distance spacing between words in the RuWordNet thesaurus is the lack of semantic connections between some members of the series, the inconsistency of their routes with the general linguistic representations fixed in the dictionaries. This fact requires a careful study of the semantic relationships of analogues in the RuWordNet thesaurus by bringing them closer together on the basis of a comprehensive analysis assuming not only a conceptual level of semantics, but also its deeper section, including various modifications of the communicative fragment.

As a result of the distances analysis between the analogue and the target word, we have determined the following main reasons for the lack of small distance between words in RuWordNet: skipping of the words meaning, referring to different semantic areas, skipping of semantic relations and skipping of concepts.

With the help of the component analysis, we identified key words considering their preservation or loss in hypo-hyperonymic relationships. The result of such approach application was the formulation not only of recommendations for RuWordNet for specific words, but also to carefully analyze and take into account the values from the explanatory dictionaries.

Application of conceptual analysis enabled to identify significant cognitive features in understanding concepts and go beyond the framework of a systematic approach to a deeper understanding of concepts and the conceptual categories of consciousness.

Acknowledgments

This research was financially supported by RFBR, grant № 18-00-01238 and by the Russian Government Program of Competitive Growth of Kazan Federal University.

References

1. A New Explanatory Dictionary of Synonyms of the Russian Language. 2nd edn. Apresyan, Y. (ed.). Yazyki russkoy kultury, Moscow (2004). (in Russian)
2. Thesaurus of Russian Language RuWordNet, https://ruwordnet.ru/ru, last accessed 2020/03/10 (in Russian)

Machine learning and text analysis in the tasks of knowledge graphs refinement and enrichment (Extended Abstract)

Victor Telnov[0000-0003-0176-5016] and Yuri Korovin[0000-0002-8399-4439]

National Research Nuclear University "MEPhI", 249040 Obninsk, Russia
telnov@bk.ru

The world of data is a place where computers rule. Supercomputers have amazing capabilities, but they often find it difficult when it comes to acquiring new knowledge and experience or existing knowledge categorization. While it's easy for a human to decide whether two or more things are related based on cognitive associations, a computer often fails to do it. The endowment of machines with common sense, as well as domain–specific knowledge in order to give them an understanding of certain problem domains, has been and remains the main goal of research in the field of artificial intelligence. While the amount of data on the WWW, as well as in corporative intranets headily grows, knowledge databases engineering still remains a challenge. This paper discusses, how the semi–automatic methods work for knowledge graphs refinement and enrichment.

A recent authoritative review of the latest achievements and current issues in the designated field of the Semantic Web is given in [1]. Our main practical contribution in this area is to develop working prototypes first, then scalable semantic web portals, which are deployed on cloud platforms and intended for use in universities educational activity. The first project [2] is related to teaching in the field of nuclear physics and nuclear power engineering. The second project [3] is related to training in computer science and programming.

The possibility of using the DL-Learner software in conjunction with the Apache Jena Reasoners in order to refine the ontologies that are designed on the basis of the SROIQ(D) description logic is shown. As a toolkit for ontologies enrichment, a software agent for the context-sensitive searching for new knowledge in the WWW has been developed. To evaluate the measure of compliance of the found content concerning a specific domain, the binary Pareto relation and Levenshtein metrics are used. The proposed semantic annotation methods allow the knowledge engineer to calculate the measure of the proximity of an arbitrary network resource about classes and objects of specific knowledge graphs. The proposed software solutions are based on cloud computing using DBaaS and PaaS service models to ensure the scalability of data warehouses and network services. Examples of using the software and technologies under discuss are given. The potential beneficiaries of solutions and technologies that are proposed in the projects mentioned above are students, professors, experts, engineers and managers, which concentrate in the specified domains.

As for related works, groups of scientists from the University of Manchester, Stanford University, University of Bari, University of Leipzig, Cambridge Semantics and

a number of other universities are focused on the issues of theory development and technology's implementation for the semantic web, description logics and machine learning. Among the publicly available working frameworks that are designed to enrich knowledge graphs with content from the WWW, the REX project should be mentioned first [4]. Also, special mention should be made on the project [5], where for the first time an attempt was made to put into practice the methods of inductive reasoning for the purpose of semantic annotation of content from the WWW. Among modern industrial solutions aimed at corporate users, special attention should be paid to Ontotext Solutions [6]. The solution categorizes unstructured information by performing knowledge graph–powered semantic analysis over the full text of the documents and applying supervised machine learning and rules that automate classification decisions. This service also analyses the text, extracts concepts, identifies topics, keywords, and important relationships, and disambiguates similar entities.

As for the use of the binary Pareto relation for multi-criteria ranking and clustering of the found network content, the use of this fruitful idea is not fundamentally innovative. For example, the Skyline software [7] has been actively using Pareto sets for working with databases for two decades. In our case, the software implementation of the Pareto optimality principle is peculiar, when a dynamically calculated dominance index allows us to categorize network content without storing it all on the computer's memory. Multi-criteria ranking of network content can be provided under the following conditions: 1) when groups of criteria are ordered by importance; 2) when the comparative importance is known only for some pairs of criteria; 3) when there is no information on the relative importance of the criteria.

It is important to continue to develop and improve tools for the intuitive perception of linked data for non-professionals. VOWL [8] is one of the modern projects for the user-oriented representation of ontologies, it proposes the visual language, which is based on a set of graphical primitives and an abstract colour scheme. LinkDaViz [9] propose a web-based implementation of workflow which guides users through the process of creating visualizations by automatically categorizing and binding data to visualization parameters. The approach is based on a heuristic analysis of the structure of the input data and a visualization model facilitating the binding between data and visualization options. SynopsViz [10] is a tool for scalable multilevel charting and visual exploration of very large RDF & Linked Data datasets.

In contrast to the above solutions, our projects [2, 3] are mainly focused on the implementation in educational activities of universities and is not limited to visualization of knowledge graphs and interactive navigation, but is aimed at the introduction of the latest semantic web technologies to the training process, taking into account the achievements in the field of uncertain reasoning. Both the results obtained and the software created are used in the real educational process at National Research Nuclear University MEPhI, and the project as a whole is focused on the practical mastering of semantic web technologies by students and professors.

Acknowledgments

The reported study was funded by the Russian Foundation for Basic Research and Government of the Kaluga Region according to the research project 19–47–400002.

References

1. d'Amato, C.: Machine Learning for the Semantic Web: Lessons learnt and next research directions. Semantic Web 11(5), 1–8 (2020) DOI: 10.3233/SW-200388.
2. Semantic educational portal. Nuclear knowledge graphs. Intelligent search agents, http://vt.obninsk.ru/x/, last accessed 2020/04/20.
3. Knowledge graphs on computer science. Intelligent search agents, http://vt.obninsk.ru/s/, last accessed 2020/04/20.
4. REX: Web-Scale Extension of RDF Knowledge Bases, http://aksw.org/Projects/REX.html, last accessed 2020/08/03.
5. d'Amato, C., Fanizzi, N., Fazzinga, B., Gottlob, G., Lukasiewicz, T.: Combining Semantic Web Search with the Power of Inductive Reasoning, http://ceur-ws.org/Vol-527/paper2.pdf, last accessed 2020/04/20.
6. Ontotext Solutions, http://www.ontotext.com/solutions/content-classification/, last accessed 2020/04/20.
7. Kalyvas, C.. Tzouramanis, T.: A Survey of Skyline Query Processing, http://arxiv.org/ftp/arxiv/papers/1704/1704.01788.pdf, last accessed 2020/08/03.
8. Schlobach, S., Janowicz, K.: Visualizing ontologies with VOWL. Semantic Web 7, 399-419 (2016) DOI: 10.3233/SW-150200
9. Thellmann, K., Galkin, M., Orlandi, F., Auer, S.: LinkDaViz – Automatic Binding of Linked Data to Visualizations. In: Proceedings of the 15th International Semantic Web Conference pp. 147-162. Bethlehem PA USA (2015).
10. Bikakis, N., Skourla, M., Papastefanatos, G.: rdf:SynopsViz – A Framework for Hierarchical Linked Data Visual Exploration and Analysis. In: Proceedings of the European Semantic Web Conference ESWC pp. 292-297. Heraklion Crete Greece (2014).

A Transformation of the RDF Mapping Language into a High-Level Data Analysis Language for Execution in a Distributed Computing Environment (Extended Abstract)

Wenfei Tang[1] and Sergey Stupnikov[2]✉[0000−0003−4720−8215]

[1] Lomonosov Moscow State University, Faculty of Computational Mathematics and Cybernetics, GSP-1, Leninskiye Gory 1-52, 119991 Moscow, Russia
[2] Institute of Informatics Problems, Federal Research Center "Computer Science and Control" of the Russian Academy of Sciences, Vavilova st. 44-2, 119333 Moscow, Russia
tangwenfei@yandex.com sstupnikov@ipiran.ru

Scientific data is an important factor in promoting scientific development and knowledge innovations. Data management methods and tools can help people organize, explore and reuse their data effectively. Obstacles on this way appear due to heterogeneity of data represented using various data models and schemas.

In order to effectively manage and reuse scientific data, a famous academic community Force11 in 2014 proposed a set of scientific data management principles called FAIR [4]. According to these principles, scientific data should be *Findable* (identifiable and described by rich metadata to be easy to find), *Accessible* by both humans and machines, *Interoperable* (should be able to be linked or integrated with other data) and *Reusable* (can be replicated and/or combined in different settings). Soon a novel interoperability architecture [5] was proposed as a reference implementation of FAIR. This architecture enhance the discovery, integration and reuse of data in repositories that lack or have incompatible APIs. It is based on RDF [2] data model and applies *RML(RDF Mapping Language)* [1] aimed to express customized mapping rules from heterogeneous data structures and serializations XML, JSON, CSV to the RDF data model.

RDF (Resource Description Framework) is essentially a data model. It provides a unified standard for describing entities and resources. Briefly speaking, it is a way and means of expressing things. An RDF specification consists of several SPO (Subject, Predicate, Object) triples.

RML is an important foundation for the implementation of FAIR principles. Modularizable RML mappings describes the structure and semantics of RDF graphs. On the other hand, RML documents are themselves RDF documents, hence RML can be published, discovered, and reused through standard web technologies and protocols. Each RML mapping describes an resource-centric graph, therefore one can make interoperable only on the data he/she is interested in. An RML mapping refers logical sources to get data from. A logical source can be a base source (any input source or base table) or a view (in case of databases).

Data from a source are mapped into RDF using *triples maps*. A triples map is a rule that maps each row of a database, each record of a CSV data source, each element of an XML data source, each object of a JSON data source, etc. into a number of RDF triples. The rule has two main parts: a *subject map* and multiple *predicate-object maps* intended to describe way of generation the respective parts of SPOs. By default, all RDF triples are placed in the default graph of the output dataset. A triples map can contain *graph maps* that place some or all of the triples into named graphs instead.

Several implementations of RML have been already developed. RMLMapper[3] runs on a normal Java environment, loads all data in memory, thus could process relatively small datasets. RMLStreamer[4] runs on Apache Flink clusters, thus could process big input files and continuous data streams.

These implementations execute RML rules to generate RDF data. This work proposes a more flexible approach with RML rules are converted into a high-level data analysis language preserving their semantics. This makes possible to generate RDF data according to RML rules in different distributed computing environments.

Hadoop[5] software framework is chosen as the primary distributed computing platform having in mind that it still holds the largest market share in big data processing. To make the RML implementation extensible and understandable the high-level language Pig Latin [3] is chosen as an implementation language. Pig Latin is an SQL-like data analysis language that can be automatically compiled into MapReduce, Tez or Spark programs to be executed in distributed Hadoop infrastructure. It allows us to focus more on data processing itself than on programming details. Besides, Pig Latin is easy to expand, and the functions of Pig Latin can be easily extended by user-defines functions (UDFs). The language is used to analyze large data sets handling structured, unstructured, and semi-structured data. Pig has a flexible, fully nested data model and allows for complex and non-atomic data types. A Pig program consists of statements, and a Pig statement takes a relation as input and produces another relation as output.

The work proposes a mapping from RML into Pig Latin as well as respective transformation framework applying *Model-driven architecture* approach.

In order to map RML into Pig, a structure of the subjects, predicates, objects, and graphs that are defined in an RML mapping should be exposed. To formalize the skeleton of mapping the *Algorithm RML2Pig* was developed. The algorithm uses the following variables: m denotes RML mapping document, dr is a directive, tm is triples map, ls is a logical source, sm is a subject map, pom is a predicate object map. Variable gms denotes a set of graph maps. Within this algorithm, directives are mapped into Pig using a semantic function $directive2Pig$. Subject maps are mapped into Pig using a semantic function $subjectMap2Pig$. Predicate-object maps are mapped into Pig using a seman-

[3] https://github.com/RMLio/rmlmapper-java
[4] https://github.com/RMLio/RMLStreamer
[5] https://hadoop.apache.org/

tic function *predicateObjectMap2Pig*. Graph maps are mapped into Pig using function *graphMap2Pig*.

Algorithm 1 RML2Pig

Input: m : *RML mapping document*
Output: pc : *set of Pig statements*

let $m = (dr, tm)$
$pc \leftarrow$ directive2Pig(dr)
for each $triplesMap \in tm$ **do**
    let $triplesMap = (ls, sm, pom)$
    $pc \leftarrow pc \cup$ logicalSource2Pig(ls)
    $pc \leftarrow pc \cup$ subjectMap2Pig(sm)
    $pc \leftarrow pc \cup$ predicateObjectMap2Pig(sm, pom)
    $gms \leftarrow gms \cup$ Graph Maps in sm and pom
end for
$pc \leftarrow pc \cup$ graphMap2Pig(gms,pc)

Table 1. The mapping of a subject map

RML	Pig

`<#VenueMapping>`
`rr:subjectMap [`
`  rr:template`
`  "http://loc.example.com/`
`  city/{$.venue[*].`
`  location.city}";`
`  rr:class schema:City ];`
	`subject =`
	`FOREACH VenueMapping_data`
	`  GENERATE R2PFORMAT('`
	`  http://loc.example.com`
	`  /city/{$.venue[*].`
	`  location.city}',$0),`
	`  'rdf:type',`
	`  'schema:City';`

To save space only *subjectMap2Pig* semantic function is illustrated here. Consider an example of subject map presented in the left column of the Table 1. It is represented in Pig by the first FOREACH statement in the right column of the table. This statement generates a target triple for each tuple of the source collection *VenueMapping_data*.

The first element of a triple is a result of a call of an UDF R2PFORMAT[6]. The second and the third elements of the triple are *rdf:type* and *schema:City* constant strings. The UDF R2PFORMAT accepts two parameters: a template chararray and a tuple. It replaces the content enclosed in curly braces within the template by the value of the second parameter and returns the result. In this

[6] https://github.com/tangwwwfei/RML2Pig/tree/master/LoadTurtle/src/main/java/r2ps/udf/pig/R2PFORMAT.java

case, the second parameter *$0* refers to the first column (*city*) of tuples from *VenueMapping_data* collection.

Table 2. Evaluation of JSON datasets transformation

JSON	1M	3M	5M	10M	50M
RML2Pig-Tez	48s	1m39s	2m21s	3m42s	20m25s
RML2Scala-Spark	1m8s	2m12s	3m31s	5m55s	28m5s
RMLStreamer	2m29s	7m9s	11m49s	24m38s	2 hours
RMLMapper	59s	-	-	-	-

The proposed mapping was implemented[7] applying Model-driven architecture approach. At first, abstract syntax metamodels of RML and Pig conforming Ecore meta-metamodel were developed. At second, RML models are extracted from RML textual representation using Xtext[8] framework. At third, the ATLAS Transformation Language (ATL)[9] was applied to implement RML into Pig mapping rules. This allows to transform any RML Ecore model into respective Pig Ecore model. ATL is also applied to convert Pig Ecore models into textual Pig code.

A comparative evaluation of the approach with its competitors RMLMapper and RMLStreamer was performed. Testing environment includes AMD Ryzen 7 3700x, 32GB RAM, and 2TB HDD with Ubuntu 20.04 operating system and Docker 19.03. All three implementations were tested using CSV, JSON and XML datasets, which have the same structure as the test data[10] of RMLStreamer. The number of records in datasets ranges from 1 million to 50 million. XML data size ranges from 232 MB to 16 GB , CSV data size ranges from 130 MB to 4.8 GB, and JSON data size ranges from 178 MB to 15 GB. To save space, results for JSON data only are presented here in Table 2.

RMLStreamer runs on Apache Flink clusters, the docker image of Apache Flink cluster is provided by the author of RMLStreamer. The Pig runs on Tez mode on the Hadoop cluster with one NameNode and two DataNodes. The Scala runs on Yarn Cluster mode on Apache Spark with one master node and two slave nodes. RMLMapper also runs on docker that is a standalone environment with OpenJDK 8.

Note that a transformation from RML into Scala were also implemented. The mapping is based on the same principles as RML into Pig mapping and share the same UDFs, so the details are omitted here.

As shown in Table 2, the running time of RMLMapper exceeds 12 hours when the number of records is larger than 1 million. So its running time is represented by the symbol "-", which means the test result is unavailable.

[7] RML2Pig Project, https://github.com/tangwwwfei/RML2Pig
[8] https://www.eclipse.org/Xtext/
[9] https://www.eclipse.org/atl/
[10] https://figshare.com/articles/Data_for_bounded_data_study/6115049

According to the evaluation results, the RMLMapper is the slowest, and has almost no ability to handle large datasets. The RML2Pig on Tez and the RML2Scala on Yarn clusters have better performance than RMLStreamer when processing XML and JSON datasets. And they have slightly better performance than RMLSteamer when processing large CSV dataset. Overall, the running time of our implementations are almost linear, and have better performance than existing implementations. It should be noted that because of the need to provide support for XPath, the XML processing largely depends on third-party libraries.

In addition, RML2Pig has passed almost all (about 95 %) basic functional tests of RML taken from RMLMapper implementation.

To conclude it should be noted that some problems have not been solved within this work yet: (1) data sources described using DCAT (Data Catalog Vocabulary) and Hydra core vocabulary for Web API description are not supported, (2) some RML features are not supported (function executions, configuration file and metadata generation), (3) only N-Quads output RDF format is supported by now. These problems are considered as a future work. Also other high level languages and computing platforms can be considered to implement RML and compare the performance.

Acknowledgement. The research is financially supported by Russian Foundation for Basic Research, projects 18-07-01434, 18-29-22096.

References

1. Dimou, A., Sande, M.V., Colpaert, P., Verborgh, R., Mannens, E., de Walle, R.V.: RML: A generic language for integrated RDF mappings of heterogeneous data. In: Bizer, C., Heath, T., Auer, S., Berners-Lee, T. (eds.) Proceedings of the Workshop on Linked Data on the Web co-located with the 23rd International World Wide Web Conference (WWW 2014), Seoul, Korea, April 8, 2014. CEUR Workshop Proceedings, vol. 1184. CEUR-WS.org (2014), http://ceur-ws.org/Vol-1184/ldow2014_paper_01.pdf
2. Klyne, G., Carroll, J.J., McBride, B.: Rdf 1.1 concepts and abstract syntax. Recommendation, W3C (2014), https://www.w3.org/TR/rdf11-concepts/
3. Olston, C., Reed, B., Srivastava, U., Kumar, R., Tomkins, A.: Pig latin: a not-so-foreign language for data processing. In: Proceedings of the 2008 ACM SIGMOD international conference on Management of data. pp. 1099–1110. ACM (2008)
4. Wilkinson, M.D., Dumontier, M., Aalbersberg, I.J., Appleton, G., Axton, M., Baak, A., Blomberg, N., Boiten, J.W., da Silva Santos, L.B., Bourne, P.E., et al.: The fair guiding principles for scientific data management and stewardship. Scientific data **3** (2016). https://doi.org/10.1038/sdata.2016.18
5. Wilkinson, M.D., Verborgh, R., da Silva Santos, L.O.B., Clark, T., Swertz, M.A., Kelpin, F.D., Gray, A.J., Schultes, E.A., van Mulligen, E.M., Ciccarese, P., et al.: Interoperability and fairness through a novel combination of web technologies. PeerJ Computer Science **3**, e110 (2017)

Navigation Tool for the Linguistic Linked Open Data Cloud in Russian and the Languages of Russia (Extended Abstract)

Konstantin Nikolaev and Alexander Kirillovich

Kazan Federal University, Kazan, Russia
Joint Supercomputer Center of the Russian Academy of Sciences, Kazan, Russia

konnikolaeff@yandex.ru, alik.kirillovich@gmail.com

Abstract. This demo is dedicated to using of LodView RDF browser for navigation on the Linguistic Linked Open Data cloud in Russian and languages of Russia. We reveal several limitations of LodView, that prevents its using for this purpose. These limitations are: 1) resolution of Cyrillic URIs; 2) Cyrillic URIs in Turtle representations of resources; 3) support for Cyrillic literals; 4) support for URIs with IDs of fragments; 5) human-readable URLs for RDF representations of resources; 6) deployment of embedded resources. We updated the LodView for fix the recovered limitations.

Keywords: LodView, RDF browser, Linked Open Data, Linguistic Linked Open Data.

LodView is one of the most popular RDF browsers in the World. This demo is dedicated to using of LodView for navigation on the Linguistic Linked Open Data cloud in Russian and languages of Russia. We reveal several issues, that prevents using of LodView for this purpose. These issues are:

Fig. 1. Error while resolving a link containing Cyrillic literals

1) Resolution of Cyrillic URLs. Datasets from the Russian-language LLOD can use URIs containing Cyrillic characters. For example, in the RuThes Cloud set, the URI used to denote the lexical unit "машина" is <http://lod.ruthes.org/resource/entry/RU-машина-n>. However, on some configurations, resolving the Cyrillic URIs in LodView leads to an error (see Figure 2). We must add support for URIs containing Cyrillic characters to LodView.

A) Encoded URIs:

```
<resource/concept/154>
 a skos:Concept;
 lemon:isReferenceOf
  <resource/sense/154-%D0%BC%D0%B0%D1%88%D0%B8%D0%BD%D0%B0-n-0>,
  <resource/sense/154-%D0%B0%D0%B2%D1%82%D0%BE%D0%BC%D0%BE%D0%
B1%D0%B8%D0%BB%D1%8C-n-0>,
 <...>.
```

B) Decoded URIs:

```
<resource/concept/154>
 a skos:Concept;
 lemon:isReferenceOf
  <resource/sense/154-машина-n-0>,   <resource/sense/154-автомобиль-n-0>,
  <...>.
```

Fig. 2. Fragment of a Turtle representation of the "Car" concept from the RuThes Cloud resource with encoded and decoded URIs

2) Cyrillic URIs in the Turtle representations of resources. In a machine-readable representation of a resource in the Turtle serialization format, Cyrillic URIs are encoded. Figure 3A shows a fragment of the Turtle representation of the "Автомобиль" concept with encoded URIs. Representing URIs in encoded form makes it difficult to perceive them. In the Turtle representation, URIs must be represented in decoded form (for example, as in Figure 3B)

3) Support for Cyrillic literals. On some configurations, literals are represented with a broken encoding in the machine-readable representation of the resource. This happens even if the dataset has Unicode encoding. Figure 4A shows a Turtle representation of the RuThes Cloud resource with a broken encoding. It is necessary to display Cyrillic literals correctly on all configurations (as in Figure 4B).

A) Broken encoding:

```
<http://lod.ruthes.org/resource/ruthes>
 a skos:ConceptScheme;
 rdfs:label
"RuThes Cloud LLOD dataset"@en.
   "РќР°Р±РѕСЂ Р»РёРЅРіРІРёСЃС‚РёС‡РµСЃРєРёС… РѕС,
РєСЂС‹С,С‹С…
   СЃРІСЏР·Р°РЅРЅС‹С… РґР°РЅРЅС‹С… Р СѓРўРµР· Cloud"@ru.
```

B) Correct encoding:

```
<http://lod.ruthes.org/resource/ruthes>
 a skos:ConceptScheme;
 rdfs:label
  "RuThes Cloud LLOD dataset"@en,
  "Набор лингвистических открытых связанных данных РуТез Cloud"@ru.
```

Fig. 3. Fragment of the Turtle representation of the RuThes Cloud resource

4) Support for URIs with fragment IDs. Datasets from the LLOD cloud can use URIs that contain Cyrillic fragment IDs. These URIs are used to indicate a resource that is closely related to some other resource. For example, in the RDF representation of the Princeton WordNet, the URI used to denote the lexical unit "cat" is <http://wordnet-rdf.princeton.edu/wn31/cat-n>, and the URI used to indicate the canonical form of this lexical unit is <http://wordnet-rdf.princeton.edu/wn31/cat-n#CanonicalForm>. In this case, the URI for the canonical form of the lexical unit "cat" consists of the URI of this lexical unit, to which the ID of the fragment "#CanonicalForm" is added.

However, LodView does not support URIs containing the IDs of the fragment. So, for example, when resolving the URI <http://wordnet-rdf.princeton.edu/wn31/cat-n#CanonicalForm>, LodView returns a representation not of the resource that is designated by this URI, but of the resource that is designated by the URI <http://wordnet-rdf.princeton.edu/wn31/cat-n>. This problem occurs because of the HTTP Protocol: when sending a request, only a fragment of the URI is sent without the fragment ID. To solve this problem, when resolving any URI, LodView must return a representation not only of the resource designated by this URI, but also of all resources whose URI is formed by combining the URI of the source resource and some fragment ID. For example, it is necessary that when resolving the URI <http://wordnet-rdf.princeton.edu/wn31/cat-n>, LodView returned not only a representation of the resource denoted by this URI, but also of the resource denoted by the URI <http://wordnet-rdf.princeton.edu/wn31/cat-n#CanonicalForm>.

5) Human-readable URLs for the RDF representation of resources. LodView implements the negation content principle in the process of resolving URIs. In accordance with this principle, when a user requests the URI of a certain resource, LodView redirects the user to the URL where the web page with human-readable

resource representation is located. When a resource URI is requested by a software agent, LodView redirects the agent to a URL with a machine representation of the resource.

URLs for representing resources as web pages have a user-friendly format: < resource URI>.html. For example, a web page for a resource with the <http://lod.ruthes.org/resource/entry/RU-машина-n> URI has the <http://lod.ruthes.org/resource/entry/RU-машина-n.html> URL. However, URLs for RDF representations of resources are not easy to read. For example, a Turtle representation for a resource with <http://lod.ruthes.org/resource/entry/RU-машина-n> has the <http://lod.ruthes.org/resource/entry/RU-машина-n?output=application%2Frdf%2Bxml> URL. Machine-readable RDF resource representations must have the form like <resource URI>.ttl, <resource URI>.n3 etc. For example, the Turtle representation of <http://lod.ruthes.org/resource/entry/RU-машина-n> resource must have the <http://lod.ruthes.org/resource/entry/RU-машина-n.ttl> URL.

6) The unfolding of the embedded resources. On a web page with a resource view, the resource properties are presented as a table. The first column contains the property name, and the second column contains its value.

If the property value is the URI of another resource, in order to find out the properties of this resource, the user must follow the link of the resource. The system must allow the user to expand embedded resources so that they can see the properties of these embedded resources without leaving the source page.

If the property value is an anonymous resource (blank node), the property value is the surrogate ID of this anonymous resource, and the description of the anonymous resource itself is located at the bottom of the page. Similarly, the system should allow the user to deploy an anonymous resource to see its properties "in place", without scrolling the page.

At the moment, we have fully implemented the corrections, resolving the issues #1-3 and #5-6, and partially the issue #4. These corrections are the main contribution of our work

ADVANCED DATA ANALYSIS METHODS

Validating psychometric survey responses* (Extended Abstract)

Alberto Mastrotto[1], Anderson Nelson[1], Dev Sharma[1], Ergeta Muca[1], Kristina Liapchin[1], Luis Losada[1], Mayur Bansal[1], and Roman S. Samarev[2,3][0000−0001−9391−2444]

[1] Columbia University, 116th St and Broadway, New York, NY 10027, USA
(am5130,an2908,ds3761,em3414,kl3127,ll3276,mb4511)@columbia.edu
https://www.columbia.edu/
[2] dotin Inc, Francisco Ln. 194, 94539, Fremont CA, USA
romans@dotin.us https://dotin.us
[3] Bauman Moscow State Technical University, ul. Baumanskaya 2-ya, 5/1, 105005, Moscow, Russia, https://bmstu.ru/en

1 Introduction

Survey responses can be a crucial data point for researchers and organization seeking to gain feedback and insight. Modern survey design incentives users to complete as many surveys as possible in order to be compensated, in some situations, users are falsifying the response, thus rendering the response invalid. Organization and researchers can reach the wrong conclusion if the user responses are largely invalid. Mouse and keyboard are most common controls available for PC users. Even now, with plenty of touch screen devices, from programmatic point of view, touch screen generates mouse related commands.

We found different directions of research of mouse tracks: mood analysis, authentication based on user specific analysis, common behaviour analysis. The work [3] demonstrates use of multimodal user identification based on keyboard and mouse activity. The authors used False Rejection Rate as a quality value and show it $\approx 3.2\%$. Main features their used for mouse analysis: traveled distance between clicks, time intervals between releasing and next pressing, and vice versa, double click values like times, time interval, distance, and similar drag-and-drop parameters. The paper [4] demonstrate mouse detection for user authentication. In the first case, only distance travelled by the mouse was used. And two hypotheses were considered: mouse speed increases with the distance travelled, mouse speed is different in different directions considerably. The key idea was to restrict the screen for mouse activity recording by a set of 9 buttons placed inside a square. The control parameters were used false acceptance and false rejection rates (FAR and FRR) with 1.53 and 5.65 maximum values respectively. In the second paper, there are additional mouse extracted features as operation frequency, silence ratio as a percent of idle time, movement time and

* To university Columbia, the project Capstone and dotin Inc. for provided data sources.

offsets, average movement time and distance, distribution of cursor positions, horizontal, vertical, tangential velocity, acceleration and jerk, slope angle and curvature. Later work [2] uses multiple classifiers (SVM, K-Nearest Neighbor and Naive Bayes classifiers) for solving the same task of user authentication and demonstrates better results $FAR \approx 0.064$ and $FRR \approx 0.576$.

2 Data

We created a survey with 16 web pages consisting of 144 questions, and collected the survey response, mouse coordinates , clicks, scrolls, and radio clicks. The survey was conducted by means the service Amazon Mechanical Turk, and we collected the country of origin and the occupation as additional data. Lastly, we also collected the dimensions of the devise the data was being taken on. The data allowed us to understand if the response survey response changed at any time, determine if the survey was being on a tablet or PC. The data highlighted that completion time varied per user. We observed instances where it would improbable to complete the survey in good faith, i.e user taking 11 seconds. As part of the data cleaning efforts we filtered the users that did not click on on all the radio buttons. Key Dataset metrics are:Total Users - 730, Average Time - 508 sec, Std Time - 313 sec, Min Time - 11 sec, 1st Quartile Time - 274 sec, 3rd Quartile Time - 667 sec, Max Time - 2856 sec.

3 Expert rules approach

From our exploratory data analysis we identified that the tracking method used to generate the mouse path dataset presented some challenges as many of the user's paths weren't fully recorded. Out of the 755 user's data, only 54 fulfilled the basic requirement of clicking the 196 radio buttons pertaining to individual questions.

Suspicious user flags:

- Anomalies by scores. We discovered that 150 of the 755 users surveyed answer at least one page of the survey with all of the same scores.
- Anomalies by time. We then proceeded to focus on the time perspective by estimating the read time that an honest user would take to read the survey and compared it with the actual completion time taken by each individual user (256 words per a minute). From our analysis, on average, a user that completed the entire survey would need 5 minutes and 30 seconds to at least read all the 196 questions, yet 33% of our surveyed users took less time than that.
- Anomalies by topic. For each topic, we aggregated questions that are either positive or negative (i.e. Tidy/Untidy) and we analyzed how users answer differently for similar questions. In our analysis, we chose a threshold for a standard deviation of 2 to identify unfocused users, consequently resulting in 33% of users answering opposite questions with similar answers, (i.e. $Tidy = 5$; $Untidy = 5$).

Combining all these three features together, we decided to select as outliers all users with a flag score 0, consequently identifying 310 users i.e. (44% of the total users).

4 LSTM based approach

In order to feed an Recurrent Neural Networks (RNN), we needed to transform our data into a sequential format that the RNN can understand. For this purpose, we created string-based tokens which identified the cardinal directions and magnitudes of a user's movements. Page changes are identified with the "pagechange" token. All of a user's movements were appended to a single tokenized list of strings. For example, a user's movements might start off as ["nw", "1", "sw", "3" .... "pagechange" "ne", "2"] accordingly to a movement direction. For memory efficiency, movements were averaged out between radio clicks.

We used Long Short-Term Memory (LSTM) as a model as they are robust against the vanishing gradient problem. Our models carried two types of parameters: token embeddings and hidden states. Weights also included those which the LSTM uses to determine how significant of an adjustment should be made for the new sequential input. We used the cross-entropy loss function, and the evaluation metric for both the language model and the classifier was Accuracy. Once the first model was trained, we replaced the final linear layer with a classification head of N x 2 dimensions, which produced a binary label where N is the input dimensions of the final hidden state from our LSTM.

In stage 1 of the language model, we trained the model on our training set and received an accuracy of $\sim 64\%$ after twenty-five epochs. Now that we have developed a model that was able to predict the next word, we removed the head which is used for language models, and replaced it with a classifier head with randomly generated parameters. Hence, we trained this head to classify the validation status of surveys. The LSTM produced a $\sim$ 90% accuracy on predicting whether a user's survey response is valid or invalid. This approach produced the highest recall.

5 HMM based approach

Our third proposed method to determine the users' authenticity in survey responses is by analyzing the sequence of user movement using a Hidden Markov Model (HMM). We converted the window aspect ratios into device types and discovered that certain users elected to take the survey on a laptop or mobile device. We solely focused on users who completed the survey using a laptop for modeling purposes, and focused on users' coordinates across the survey duration and discovered that there's a lot of noise in the movements. To run an effective model, we converted the coordinates into discrete observations representing cardinal directions very similar to what we have used for LSTM approach.

We recognize that users are navigating through survey pages, so we use the coordinates of the next button to estimate when each user moves to the next

page. After analyzing each survey page, we realized that each user has a unique layout and the mouse path that users exhibit varies. Furthermore, considering that the number of mouse movement records varies per page, we decided to analyze the first 200 observations per user. We also removed the users that took the survey multiple times. After multiple attempts those users have become accustomed to the survey design and movement would be based on memory.

Only 66 users met the defined criteria for further analysis in this approach. We trained the HMM using the Baum-Welch algorithm to estimate the transition matrix, state distribution, and output distribution. We train the algorithm to recognize the patterns in each page and apply the forward algorithm to calculate the observation log probability of each observed user sequence per page. A low log probability is interpreted as having a less likely occurrence.

We scale each observation and apply an isolation forest to identify those suspicious users. Out of the 66 users, 11%, or 7 users were labeled as suspicious.

The accuracy of the HMM is dependent on the validity of the assumptions, and the quality of the data [1], [5]. We therefore identify the assumptions and limitations of this approach.

- The captured data doesn't distinguish when users are using their mouse to complete the survey vs browsing the internet.
- The model assumes that the majority of users are completing the survey in good faith. If most users are falsely completing the survey, then the users that are attempting to complete the survey in good faith will be flagged.
- The page labels were estimated using the coordinates of the next button on each page. Those labels represent our best estimate and may not truly reflect when the user page changes.

References

1. Elbahi, A., Omri, M.N., Mahjoub, M.A., Garrouch, K.: Mouse movement and probabilistic graphical models based e-learning activity recognition improvement possibilistic model. Arabian Journal for Science and Engineering **41**(8), 2847–2862 (2016). https://doi.org/10.1007/s13369-016-2025-6, https://doi.org/10.1007/s13369-016-2025-6
2. Karim, M., Heickal, H., Hasanuzzaman, M.: User authentication from mouse movement data using multiple classifiers. In: Proceedings of the 9th International Conference on Machine Learning and Computing. p. 122–127. ICMLC 2017, Association for Computing Machinery, New York, NY, USA (2017). https://doi.org/10.1145/3055635.3056620, https://doi.org/10.1145/3055635.3056620
3. Motwani, A., Jain, R., Sondhi, J.: A multimodal behavioral biometric technique for user identification using mouse and keystroke dynamics. International Journal of Computer Applications **111**, 15–20 (02 2015). https://doi.org/10.5120/19558-1307
4. Singh, S., Arya, K.V.: Mouse interaction based authentication system by classifying the distance travelled by the mouse. International Journal of Computer Applications **17** (03 2011). https://doi.org/10.5120/2181-2752
5. Stamp, M.: Introduction to Machine Learning with Applications in Information Security. Chapman & Hall/CRC, 1st edn. (2017)

Comparison of Two Approaches to Recommender Systems with Anonymous Purchase Data* (Extended Abstract)

Yuri Zhuravlev[1], Alexander Dokukin[1], Oleg Senko[1], Dmitry Stefanovsky[2], Ivan Saenko[3], and Nikolay Korolev[3]

[1] FRC CSC RAS, Moscow, Russia
zhur@ccas.ru, dalex@ccas.ru, senkoov@mail.ru
[2] IRIAS, Moscow, Russia
dstefanovskiy@gmail.com
[3] Moscow State University, Moscow, Russia
i.a.saenko@mail.ru
[4] Moscow State University, Moscow, Russia
nikolay.korolev.s@gmail.com

Abstract. A new approach for recommender systems design is discussed. The considered system should rely only on anonymous receipts' data and information about products currently bought by a customer. The preference rating for an arbitrary product is calculated as a classification result of a combined feature description of a product that currently is being bought and products that have been bought previously by the same customer. Two different approaches aimed to calculate such descriptions are proposed. The methods were compared with two other techniques in experiments with real retail data, that is estimating preference rating simply as a product sales rate and using association rules. It was shown by experiments that proposed methods outperform two latter ones in terms of areas under ROC curves.

Keywords: Recommender system · Machine learning · Feature extraction · Gradient boosting.

The purpose of this article is to study the effective use of data to personalize offers to retail customers, mainly with online purchases. In general, the model of the relationship between the retailer and the buyer is as follows: the appearance of the client is considered consistent, so the seller offers each buyer a range of products, then the client decides whether to make a purchase. The retailer may encounter restrictions in terms of display or capacity, which limit the number of products in the offer. The retailers goal is to maximize the expected total revenue for the sales season. Recommender systems are a popular tool aimed to give a

* This work was supported in part by the Russian Foundation for Basic Research, projects no. 18-29-03151, 18-01-00557.

customer an advice which good in the best way corresponds to his/her demands [1]. Many techniques can be used to implement one. Context based systems use some supplementary information about customers or goods. However such information is often hard to achieve. Another approach employs information about customers preferences expressed by them in one way or another. In the latter case some very efficient mathematical tools involving matrix decomposition can be used. But getting clients preference data is associated with additional costs in offline shopping. Finally, recommendations can be based on digital traces left by anonymous customers, i.e. the set of customers receipts registered up to a certain point. Additionally in the simplest case all identification is based on a set of goods being bought. Segmentation or clustering is the key to effective personalization and identifying preferences.

We illustrate the practical value of a clustering policy in real conditions using a dataset from a major Russian retailer. The data set consists of roughly ten thousand cosmetic and related goods purchased in different combinations in about one hundred thousand transactions over two months period. We compare the effectiveness of the proposed policy with a data-intensive policy that ignores any potential similarity of preferences in different profiles and, thus, evaluates the product preferences for each profile individually.

Association rule is a way of measuring consequence like relationships between objects [2]. In this case the relation between a product X being bought, i.e. $Z_0(X) = 1$, and a product Y to be recommended. Here, each receipt Z is described as a binary vector of length N corresponding to a total number of products. New customers recommendations are made on the basis of the previously collected receipt data and the currently performed transaction Z_0 which can be described in the same binary form.

Lets denote $S(\{X\})$ and $S(\{Y\})$ the subsets of all receipts S containing product X or Y correspondingly, whereas $S(\{X, Y\})$ will denote a subset containing both goods. The support of X is then defined as $Sup(X) = \frac{|S(\{X\})|}{|S|}$, whereas confidence of the X, Y pair is $Conf(X, Y) = \frac{|S(\{X,Y\})|}{|S(\{X\})|}$. Pairs of objects with large enough values of both criteria form association rules which can be used to estimate probability of buying product Y subject to product X purchase. The conclusion may be made based on a single best association rule or by their ensemble.

Another approach, namely the baseline frequency based algorithm, calculates receipt Z_j owner preference ratings for item Y as support value only, i.e.

$$A_F(Y, Z_j) = Sup(Y)\,.$$

When using clustering techniques the set of binary receipt vectors is divided into several groups or clusters in which the digital traces are considered close to each other in terms of a selected metric. Then the Y products preference rating can be calculated by frequencies present in the cluster containing the Z_0 trace.

Clustering methods are used in recommendation systems to select groups of customers with similar preference profiles [3]. Here we suggest another technique where clustering is used to select groups of complementary products. The derived

set of clusters is further used to generate multidimentional feature description of products. Such descriptions allow effective application of machine learning tools.

The authors of the present research have already shown that agglomerative hierarchical grouping method applied to the described binary data produces well interpreted set of product clusters [4]. At that the chi-squared metric [5].

Let we have L non-intersecting clusters $C_1, \ldots, C_L$ in the S_T set. The distance of product Y to the i-th cluster is calculated as

$$P(Y, C_i) = \frac{1}{|C_i|} \sum_{X \in C_i} \rho(G(Y), G(X)) .$$

Vector $P(Y) = [P(Y, C_1), \ldots, P(Y, C_L)]$ can serve as a good feature description of the product Y since it is continuous and it reflects customers interest to the product in terms of his interest to different clusters of products.

In our studies, we consider two methods for calculating customer preference estimates based on clustering. In the first method, the preference of the product for the owner of receipt Z is calculated by concatenating descriptions of r top products from Z and description of evaluated product Y. In the second method, the preference of the product for the owner of receipt Z is calculated by averaging descriptions of all products from Z combined with the description of the evaluated product Y.

The combined training sample was then used to train different two-class machine learning methods [6] A_{ML1} including logistic regression, support vector machines, decision forests with gradient boosting. Their performance was estimated with the multifold cross-validation with the class 1 gap as preference rating for a product. Given the restrictions of the current article we limit the show of results to a single case.

The ROC curve [7] for the second proposed method A_{ML2} (using LightGBM [8] method as a machine learning algorithm) is shown at the figure (1) together with ROC curves for A_{AR} (associative rules) and A_F (frequency based). In the legend "boostings" stands for A_{ML2}. It is seen from figure (1) that the ROC curve for A_{ML2} runs noticeably higher the ROC curve for A_F at interval for FPR from 0 up to 0.6. At that the ROC curve for A_{ML2} practically coincides with ROC curve for A_{AR} at interval from 0 to 0.07.

A new method has been developed for estimating customer preferences by the anonymous cash receipts data. The experiments indicate the prospects of the proposed approach.

Firstly, the effectiveness of the proposed method turned out to be slightly higher than the effectiveness of reference methods. Evaluation was performed by ROC AUC.

Secondly, it is important to mention that though the ROC AUC values of the proposed method the frequency based algorithm are quite close the experiments showed significant differences between their recommendations in terms of goods. The machine learning algorithm suggests rarer products which may be advantageous for the shop owner. Also, the correlation value indicate that method ensembles might provide some improvement in future research.

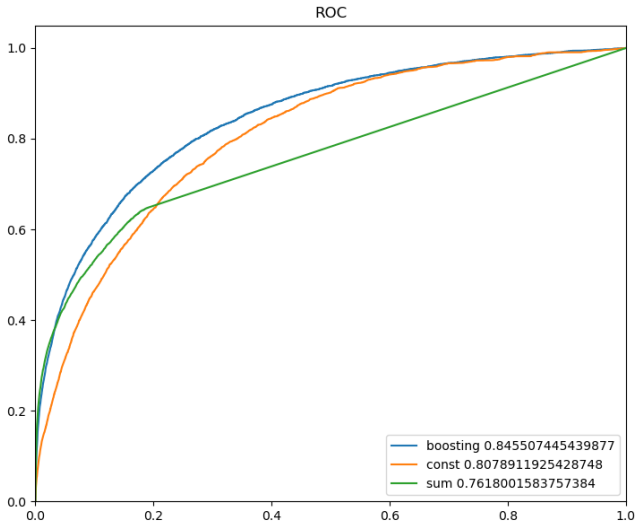


Fig. 1. ROC curve comparison for $r = 4$

References

1. Sohlberg, H.: Recommending new items to customers - a comparison between Collaborative Filtering and Association Rule Mining. Masters Thesis. Stockholm: KTH Royal institute of technology.school of computer science and communication (CSC) (2015)
2. Sun, X., Kong, F., Chen, H.: Using Quantitative Association Rules in Collaborative Filtering. In: Fan W., Wu Z., Yang J. (eds) Advances in Web-Age Information Management. WAIM 2005. Lecture Notes in Computer Science **3739**, (2005).
3. West, J.D., Wesley-Smith, I., Bergstrom, C.T.: A recommendation system based on hierarchical clustering of an article-level citation network. IEEE Transactions on Big Data **2**(2)113–123 (2020)
4. Zhuravlev, Yu., Dokukin, A., Senko, O., Stefanovskiy, D.: Use of Clasterization Technique to Highlight Groups of Related Goods by Digital Traces in Retail Trade. Proceedings of 9th International Conference on Advanced Computer Information Technologies ACIT-2019, 84–88 (2019)
5. Choi, S.-S., Cha, S.-H., Tappert, C. C.: A survey of binary similarity and distance measures. Journal of Systemics, Cybernetics and Informatics **8**(1), 43–48 (2010)
6. Hastie, T., Tibshirani, R., Friedman, J.: The Elements of Statistical Learning: Data Mining, Inference, and Prediction. Springer Science & Business Media (2013)
7. Fawcett, T.: An introduction to ROC analysis. Pattern Recognition Letters **27**, 861–874 (2006)
8. Guolin Ke, Qi Meng, Thomas Finley, Taifeng Wang, Wei Chen, Weidong Ma, Qiwei Ye, Tie-Yan Liu: LightGBM: A Highly Efficient Gradient Boosting Decision Tree. Advances in Neural Information Processing Systems 30 (NIPS 2017), 3149–3157 (2017).

The Study of the Sequential Inclusion of Paths in the Analysis of Program Code for the Task of Selecting Input Test Data (Extended Abstract)

K. E. Serdyukov[1][0000-0003-4162-1295] and T. V. Avdeenko [2][0000-0002-8614-5934]

[1] Novosibirsk State Technical University, K. Marks avenue 20, 630073 Novosibirsk, Russia

Abstract. The article proposes the results of a study evaluating one and many paths of program code, the work is not finished yet. To solve the problem of generating data sets, it is proposed to use a genetic algorithm with various methods for determining the complexity of program code. A new method is proposed for determining code complexity based on changing weights of nested operations. The article presents the results and comparison of the generated input test data for the passage along the critical path. For each metric considered in the article, conclusions are presented to identify features depending on the selected data.

Keywords: genetic algorithm, test data, control flow graph, fitness function, testing process, program code, optimization method, test data generation, code coverage

1 Problem description

We will not describe the basic idea of a genetic algorithm here, since it is well known. We indicate only its adaptation as applied to the problem being solved. In this case, the chromosomes are the data variants (values f variables) received at the program input. In this paper, we assume that the data is numerical. The purpose of the algorithm is to select such a multitude of test data variants that would provide the maximum coverage of the code, that is, to pass through as many program operations as possible. This goal is realized through the formulation of the fitness function of a certain kind, which expresses how well a particular data set is adapted for mutual coexistence with other individuals (data variants) in a given generation (a set of data variants that should generate paths that are as different from each other as possible in the code graph).

The task of finding input test data consists of three subtasks:

1. Search for input data for passing along one of the most complex code paths. Difficulty is determined by the metric chosen to evaluate the code;
2. The exclusion or reduction of the weights of operations on the path for which the data were selected, based on the fitness function for other paths;

3. Generation of input test data for many paths of program code.

The limit on the number of sets of input data is established after the development phase and will allow you to concentrate on certain paths. The method is shown in more detail in Figure 1. Inside the general method, there are operations for generating data for one path, checking for reaching restrictions and excluding operations from further selection.

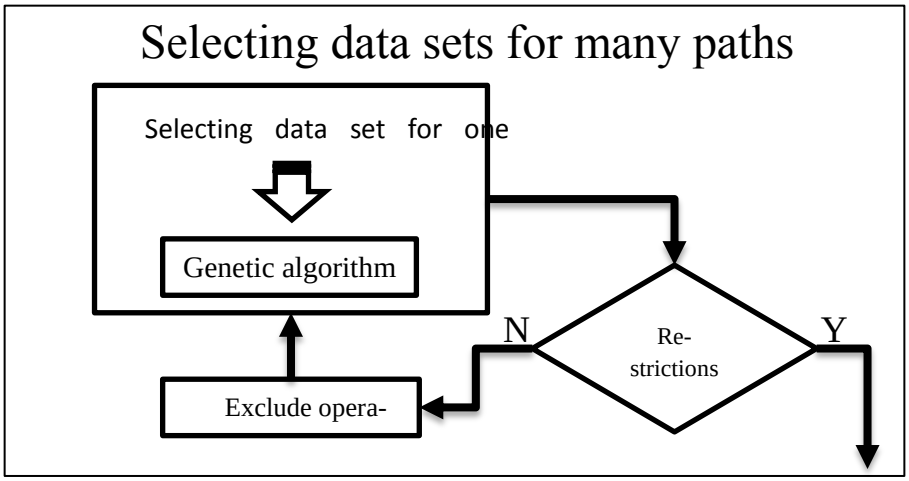


Figure 1 – Cyclic method of generating test data sets for many paths

The whole algorithm is performed cyclically - the procedure for searching for input data for one branch is started, after which operations in this branch are excluded from further calculations and the data search for one branch is started again.

2 Selection of test data sets for one and many paths

To select data for many paths, firstly show how the method works for one complex path. Modified SLOC metric is used for assessing complexity, the description of which is given in section 3.

In the algorithm, the first population is formed randomly. Certain settings were made for testing – each population contains 100 chromosomes; the total number of populations also equals 100. This will make it possible to form a sufficient number of different variants.

Table 1 presents 4 runs of the method for the program on the Figure 1 with the first random population, two middle populations and the final one, from which the first chromosome is taken and counted as the final generated data set. For convenience, only the 5 "best" chromosomes in each population will be shown.

Table 1. Different variants of the test data set for one path

Population	Run 1	Run 2	Run 3	Run 4
0	1: 78, 23, 35	1: 97, 3, 6	1: 92, 97, 28	1: 15, 67, 26
	2: 62, 36, 95	2: 82, 77, 64	2: 38, 66, 52	2: 32, 27, 83
	3: 52, 35, 27	3: 24, 47, 57	3: 63, 76, 64	3: 37, 52, 64
	4: 17, 77, 73	4: 90, 13, 82	4: 7, 24, 56	4: 70, 49, 64
	5: 75, 9, 96	5: 81, 69, 24	5: 57, 48, 8	5: 67, 29, 94
20	1: 95, 64, 54	1: 97, 80, 4	1: 99, 13, 10	1: 99, 71, 45
	2: 95, 64, 29	2: 97, 80, 53	2: 99, 13, 11	2: 99, 71, 15
	3: 95, 64, 54	3: 97, 80, 28	3: 99, 13, 11	3: 99, 71, 3
50	1: 95, 64, 54	1: 97, 80, 29	1: 99, 13, 10	1: 99, 71, 60
	2: 95, 64, 29	2: 97, 80, 4	2: 99, 13, 11	2: 99, 71, 3
	3: 95, 64, 54	3: 97, 80, 53	3: 99, 13, 11	3: 99, 71, 3
Final	*1: 95, 64, 54*	*1: 97, 80, 4*	*1: 99, 13, 10*	*1: 99, 71, 60*
	2: 95, 64, 29	2: 97, 80, 29	2: 99, 13, 11	2: 99, 71, 45

As it can be seeing, at least 2 different final sets of test data were formed in each of the variants, in which the operations in the considered program code will have the greatest weight. In addition, there is certain patterns in the results - the first value is always the maximum (random values are limited to a maximum of 100 to increase convergence), the second value is less than the first, but more than the third.

For this program code test data sets obtained in the latest generation can be used as initial data sets to testing.

In order to select data for many paths of the program code, operations that has already been selected could be sequentially excluded. Therefore, for the next test, another program is used with 70 operations and much more conditions and cycles.

To confirm the operability of the method, some of the paths of the program code are preliminarily enclosed in conditions that cannot be achieved. As a result, 8 operations cannot be executed, so code coverage cannot be 100%. Table 2 presents the results of running the method and it can be clearly seen that these paths are not achieved when selecting data.

Table 2. Test data sets for many program paths

Iteration	Data Set	Function	Code coverage
1	(77, 18, 99)	164 100	41% (29/70)
2	(60, 41, 61)	26 400	53% (37/70)
3	(5, 8, 56)	18 882	74% (52/70)
4	(40, 30, 55)	9 900	79% (55/70)
5	(99, 73, 73)	12 000	86% (60/70)
6	(84, 22, 64)	5 000	**89% (62/70)**
7	(48, 72, 44)	0	89% (62/70)
8	(36, 2, 36)	0	89% (62/70)

For the first 6 iterations of operation, the algorithm went over all possible code paths. In total, code coverage was 89%, with the exception of those operations that are on impassable paths.

Due to the fact that the data is initially randomly selected, the chance of reaching a certain path can be quite small due to possible restrictions. This problem can be solved in several ways, for example, to limit the boundaries of the selected values due to the introduction of domains or increase the chance of mutation. An increase in the chance of a mutation will allow branching out due to a greater number of combinations, but may adversely affect the overall convergence rate of the method.

3 Conclusion

The selection of input test data is an important task to ensure high quality testing. This is due to the fact that what parts of the code the tester can verify depends on the input data. Therefore, ensuring maximum coverage of the code is one of the most important goals when selecting test data.

In the proposed method, there are two tools for providing maximum coverage. Different metrics have different functionality and can lead to the selection of data for different paths. Each metric has a different way for determining fitness functions and can be used to fulfill different testing requirements.

There is still much room for further research in this area. Starting from the implementation of other common metrics and ending with the development of a hybrid metric that can be flexibly configured to solve specific problems.

Achieving maximum coverage is ensured through the use of a genetic algorithm. Despite the fact that the method works quite efficiently, there is a lot of room for improvement. The selection of test data for one path quickly came to one solution and most of the generation wasted. Therefore, in this part, the introduction of additional restrictions is necessary, since the efficiency of the algorithm as a whole directly depends on this.

When selecting data for multiple paths, it is also important to propose additional methods. In particular, it is planned to introduce an adaptive mutation, which depends on how high a level of coverage has already been achieved and whether data with a zero value of the fitness function , but not with a maximum coverage, have been obtained.

Acknowledgments

The reported study was funded by RFBR, project number 19-37-90156
The research is supported by Ministry of Science and Higher Education of Russian Federation (project No. FSUN-2020-0009).

An Information System for Inorganic Substances Physical Properties Prediction Based on Machine Learning Methods (Extended Abstract)

V.A. Dudarev[1,2][0000-0001-7243-9096], N.N. Kiselyova[1][0000-0002-3583-0704], A.V. Stolyarenko[1], A.A. Dokukin[3], O.V. Senko[3], V.V. Ryazanov[3], E.A. Vashchenko[4], M.A. Vitushko[4] and V.S. Pereverzev-Orlov[4]

[1] A.A. Baikov Institute of Metallurgy and Materials Science of RAS (IMET RAS), Moscow, 119334, Russia
[2] National Research University Higher School of Economics, Moscow, 109028, Russia
[3] Federal Research Center "Computer Science and Control" of RAS (FRC CSC RAS), Moscow, 119333, Russia
[4] A.A. Kharkevich Institute for Information Transmission Problems of RAS (IITP RAS), Moscow, 127051, Russia
kis@imet.ac.ru

Abstract. ParIS (Parameters of Inorganic Substances) system was developed for predicting inorganic substances physical properties. It is based on the use of machine learning methods to find the relationships between inorganic substances parameters and the properties of chemical elements. The main components of the system are an integrated database system on inorganic substances and materials properties, a subsystem of machine learning and prediction results analysis, a knowledge base and a prediction database. The machine learning subsystem includes programs based on the algorithms developed by the authors of this paper and the algorithms included in the scikit-learn package. The results of the ParIS system application are illustrated by an example of predicting chalcospinels crystal lattice parameter. To get prediction results, only the properties of chemical elements included in the composition of not yet synthesized chalcospinels were used. Moreover, the prediction accuracy was within ± 0.1 Å.

Keywords: Machine Learning, Databases, Prediction of Inorganic Substances Physical Properties.

1 Introduction

Machine learning methods are widely used in chemistry. Object classification and qualitative (categorical) characteristics prediction tasks are among the most successfully solved problems. However, the vast majority of problems are associated with the quantitative objects' characteristics prediction. The use of regularization, various methods of the most important features selecting, filtering of erroneous outliers, in

many cases allows to circumvent some of these limitations. However, the task of developing combined methods that would overcome most of the limitations in solving the problems of reconstructing multivariate regression cannot be considered as completely solved.

2 Selection of Machine Learning Methods for Prediction of Inorganic Compounds Physical Properties

Different ML-methods have their own strengths and weaknesses. E.g. the *support vector regression (SVR)* algorithm [1] is very sensitive to data outliers while chemical problems, as a rule, contain erroneous and out-of-date experimental values. The overfitting problem can be solved by means of regularization. *Artificial neural networks* learning can be used both for calculating functions of qualitative and quantitative parameters. The method disadvantages include the lack of modeling transparency, which does not allow a physical interpretation of the results obtained, the complexity of choosing a network architecture, high requirements for measurement errors, the complexity of choosing a learning algorithm, and high resource consumption of neural networks learning process.

The creation of combined algorithms is one of the promising modern trends in the development of methods for predicting quantitative properties. This approach makes it possible to compensate the shortcomings of some algorithms at the expense of the advantages of others and is aimed at improving the prediction accuracy of quantitative parameters, as one of the main criteria for methods effectiveness. Possible approaches are a combination of classification algorithms, an elastic network, combinations of SVR and multidimensional regression, etc. The following algorithms and programs that implement this approach are included in the system for predicting inorganic compounds physical properties:

- Locally Optimal Convex Combinations (LOCC) [2];
- Gluing Classifications for Regression (GCR) [3];
- Soft Voting Clique-Based Solvers [4, 5];
- Algorithms from scikit-learn software package [6].

3 System structure for inorganic substances physical properties prediction

The information base for searching the dependences of inorganic substances parameters on the properties of components in the ParIS system is the integrated database system on properties of inorganic substances and materials (DB PISM) that we created [7]. It virtually unites seven databases developed in Russia and Japan, and contains information on tens of thousands of inorganic substances and materials.

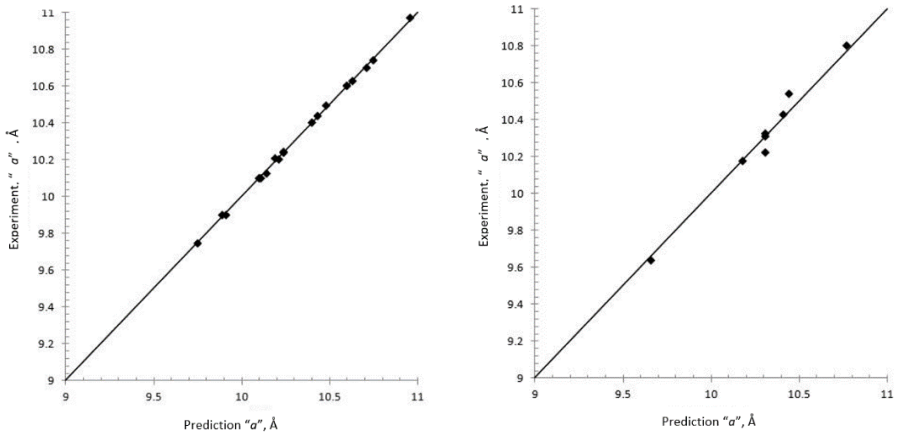
The machine learning subsystem includes three components:

- a subsystem for searching for dependencies between the substances' properties and components parameters (a machine learning subsystem) based on the programs we developed and the scikit-learn software package;
- prediction subsystem using the found dependencies;
- a subsystem for estimating prediction accuracy, which allows one to estimate the mean absolute and mean square errors (with cross-validation), the R^2 determination coefficient, etc., as well as construct a diagram of deviations of the calculated parameter values from the experimental ones for the substances, information about which was used for machine learning.

The dependencies obtained during machine learning process are entered into the knowledge base. They can be used to predict the parameters of substances not yet obtained in certain composition.

4 The ParIS System Application for Inorganic Compounds Physical Properties Prediction

The developed system was used to predict the crystal lattice parameter of not yet obtained chalcospinels – promising materials for creating magneto-optical memory elements and sensors. Predicting of crystal lattice parameters of compounds is of great interest for both chemical research and materials science investigations.



a b

Fig. 1. Comparison of the values predicted using multilevel predicting (random forest + elastic network) of the chalcospinel crystal lattice parameter with experimental values for $A^{I}B^{III}C^{IV}X_4$ (a) and $A^{II}B^{III}C^{III}S_4$ (b) compositions.

In fig. 1, the prediction results of the crystal lattice parameter for known chalcospinels using a multilevel approach, which is a combination of Random Forest and Elastic

Net machine learning methods, are presented in graphical form. It should be noted that such a multilevel method provided the smallest prediction errors.

5 Conclusion

The ParIS system was developed for inorganic substances physical properties prediction. It allows a search for the relationships between inorganic compounds physical properties and chemical elements parameters by means of machine learning analysis of information contained in databases on inorganic substances properties. The main components of the system are an integrated system of databases on inorganic substances and materials properties developed in Russia and abroad, a machine learning-based data analysis subsystem for making predictions and a knowledge base for prediction results. The ML-subsystem includes programs based on the original algorithms developed by the authors of this paper together with methods implemented in the scikit-learn package. Using the developed system, "*a"* crystal lattice parameter values have been successfully predicted for not yet obtained chalcospinels with $ABCX_4$ composition (A, B and C are various chemical elements, and X is S or Se). During prediction chemical elements properties values were used only. Moreover, the prediction accuracy was ±0.1 Å. Thus, it is shown that the original multilevel method developed by the authors provided the smallest predicting errors.

This work was supported in part by the Russian Foundation for Basic Research, project nos. 18-07-00080 and 20-01-00609. The study was carried out as part of the state assignment (project no. 075-00947-20-00).

References

1. Smola, A.J., Schölkopf, B.: A tutorial on support vector regression. Statistics and Computing. 14(3) (2004).
2. Senko, O. V., Dokukin, A. A., Kiselyova, N. N., Khomutov, N. Yu.: Two-Stage Method for Constructing Linear Regressions Using Optimal Convex Combinations. Doklady Mathematics. 97(2), 113-114 (2018).
3. Lukanin, A.A., Ryazanov, V.V., Kiselyova, N.N.: Prediction Based on the Solution of the Set of Classification Problems with Supervised Learning and Degrees of Membership. Pattern Recognition and Image Analysis. 30(1), 63-69 (2020).
4. Vaschenko, E., Vitushko, M., Dudarev, V., et al.: On the possibility of predicting the parameter values of multicomponent inorganic compounds. Information processes. 19(4), 415-432 (2019). (in russian)
5. Vaschenko, E., Vitushko, M., Pereverzev-Orlov, V.: Potentials of Learning on the Basis of Partner System. Pattern Recognition and Image Analysis. 14(1), 84-91 (2004).
6. Pedregosa, F., Varoquaux, G., Gramfort, A., et al.: Scikit-learn: Machine learning in Python. J. Machine Learning Research 12(Oct.), 2825-2830 (2011).
7. Kiselyova, N.N., Dudarev,V.A., Stolyarenko, A.V.: Integrated system of databases on the properties of inorganic substances and materials. High Temperature, 54(2), 215-222 (2016).

Расширяемая Система для Многокритериального Поиска Выбросов в Данных (Расширенные тезисы)

В. Д. Динеев [0000-0002-0218-0338], В. А. Дударев [0000-0001-7243-9096]

Национальный исследовательский университет «Высшая школа экономики», 109028, Россия
vddineev@edu.hse.ru

Аннотация. Статья посвящена созданию удобного расширяемого инструмента для поиска выбросов в данных. Разработанная информационная система базируется на использовании методов статистического анализа и машинного обучения для поиска выбросов в данных и применении Веб-технологий и микросервисной архитектуры для реализации пользовательского интерфейса и расширяемости системы. В результате разработки получен программный инструмент, способный анализировать многомерные числовые данные и находить в них выбросы с помощью набора настраиваемых методов анализа с возможностью голосования алгоритмов. Новые алгоритмы добавляются в состав системы в качестве микросервисов, взаимодействующих с родительским сервисом. Доступ конечных пользователей к системе может осуществляться через Веб-приложение с помощью Веб-браузера. Разработанный инструмент может найти применение при анализе данных и результатов экспериментов, в которых потенциально могут содержаться ошибки, добавляя необходимую степень автоматизации для эксперта, анализирующего корректность данных.

Ключевые слова: поиск выбросов в данных, анализ данных, очистка данных, Веб-технологии.

1 Введение

Поиск выбросов в данных [1] на сегодняшний день используется во многих областях: от обнаружения неисправностей в показаниях приборов до обнаружения подозрительной активности клиентов банков.

Существующие наиболее популярные инструменты поиска выбросов в основном представляют собой платные приложения (например STATISTICA [2]), которые устанавливаются на персональный компьютер. Несмотря на обширный функционал, такие приложения ограничивают пользователя в выборе технических и программных средств для использования данных решений, при этом требуя от пользователя покупки дополнительного, возможно, не нужного ему функционала. Существуют и бесплатные open-source решения, которые, на наш

взгляд, являются пригодными только для потребителей, хорошо разбирающихся в ИТ [3]. Часто отмечается разработка авторских методов, нацеленных на учет специфики данных в конкретной предметной области и последующее написание соответствующих проприетарных систем [4]. В связи с этим возникает актуальность разработки открытого модульного приложения, которое будет поддерживаться большинством существующих на данный момент операционных систем, распространяться бесплатно и доступно пользователям из прикладных областей без специальной математической подготовки или знаний в ИТ.

Для поддержки кроссплатформенности, оптимальным подходом является предоставление необходимой функциональности в виде Веб-сервиса (для программных средств) и Веб-приложения (для конечных пользователей) с доступом из сети интернет, для чего в работе были использованы Веб-технологии [5].

Поскольку существует большое количество различных алгоритмов поиска выбросов и процедур коллективного принятия решений, было решено применить микросервисную архитектуру [6].

2 Методы

2.1 Статистические критерии

Для нахождения выбросов на одномерных данных в работе использовались распространенные статистические критерии, такие как: критерий Граббса, Шовене, правило трёх сигм и тест Диксона.

2.2 Методы Машинного Обучения

Для многомерных данных лучше всего подходят алгроитмы машинного обучения. Использовались следующие алгоритмы – IsolationForest[11], Local Outlier Factor[12] и OneClassSVM[13].

2.3 Процедуры Коллективного Принятия Решения

Для получения более достоверного результата при использовании нескольких алгоритмов поиска выбросов используются такие комбинации результатов, как среднее значение и голосование по большинству.

3 Описание Разработанной Системы

Разработанный инструмент построен на основе архитектуры микросервисов (рис. 1). Каждый модуль реализуется в виде самостоятельного микросервиса, доступного через REST API и выполняющего вычисления над заданными оболочкой входными данными.

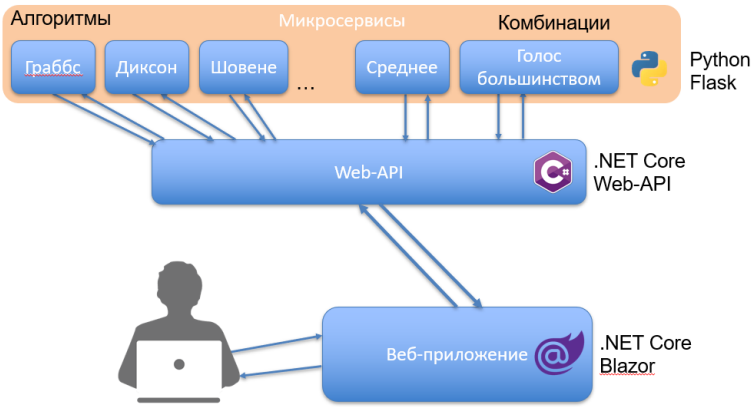


Рисунок 1. Архитектура системы.

4 Применение системы на примере поиска выбросов в обучающей выборке задачи из неорганической химии

Приведем пример использования разработанной системы для решения задачи поиска выбросов (ошибок) в обучающей выборке для прогнозирования возможности образования и типа кристаллической структуры соединений состава $A^{2+}_2 B^{+3}C^{+5}O_6$. Обучающая выборка состоит из 551 прецедента, характеризующегося вектором размера 105 (или точкой в 105-мерном пространстве признаков), отнесенного к одному из 11 классов (от отсутствия соединения до образования соединений со специфическими кристаллическими структурами).

При анализе обучающей выборки на предмет выбросов мы использовали усреднение бинарного ответа хорошо зарекомендовавших себя алгоритмов Isolation Forest и Local Outlier Factor.

В исследуемой задаче найдено пять “подозрительных” объектов-соединений, класс которых, по мнению всего коллектива методов, является ошибочным и подлежит тщательной проверке специалистом-предметником (на предмет возможной ошибки в классификации объектов обучающей выборки): Ba_2LaPuO_6 – скорее всего не имеет кристаллическую структуру K_3FeF_6-I, пр.гр. Fm3(-)m, Z=4; Ba_2RhTaO_6 и Ba_2RhUO_6 - не относятся к типу $BaTiO_3$(V), пр.гр. $P6_3/mmc$, Z=6; Sr_2RhTaO_6 – не относится к типу I4/m; а отсутствие соединения для CaO-Rh_2O_3-Ir_2O_5 указано ошибочно. Таким образом, использование программы позволяет существенно сузить количество объектов для проверки на выбросы (нужно проверить 5 из 551).

5 Заключение

В результате было разработано расширяемое Веб-приложение, основанное на микросервисной архитектуре (исходный код доступен по адресу https://github.com/dineev-vd/MultiCriteriaOutlierSearch, развернуто в тестовом режиме по адресу http://outliers.imet-db.ru/outliers), которое способно анализировать данные на наличие в них выбросов.

Работа выполнена при частичной финансовой поддержке РФФИ, проект 18-07-00080.

Литературные источники

1. Zimek A., Schubert E. Outlier Detection // Encyclopedia of Database Systems. — Springer New York (2017). DOI: https://doi.org/10.1007/978-1-4899-7993-3_80719-1
2. Боровиков В.: STATISTICA. Искусство анализа данных на компьютере: Для профессионалов: 2-е изд. СПб: Питер, (2003).
3. Flach, M., Gans, F., Brenning, A., Denzler, J., Reichstein, M., Rodner, E., Bathiany, S., Bodesheim, P., Guanche, Y., Sippel, S., and Mahecha, M. D.: Multivariate anomaly detection for Earth observations: a comparison of algorithms and feature extraction techniques, Earth Syst. Dynam., 8, 677–696, https://doi.org/10.5194/esd-8-677-2017, 2017.
4. Ожерельев И.С., Сенько О. В., Киселева Н.Н. Метод поиска выпадающих объектов с использованием параметров неустойчивости обучения // Системы и средства информатики. - 2019. - Т.29. - N.2. - С.122-134.
5. Чамберс Д., Пэкетт Д., Тиммс С.: ASP.NET Core. Разработка приложений. – СПб.: Питер (2018).
6. Ричардсон К.: Микросервисы. Паттерны разработки и рефакторинга. – СПб.: Питер (2019).
7. Лемешко Б. Ю., Лемешко С. Б.: Расширение области применения критериев типа Граббса, используемых при отбраковке аномальных измерений // Измерительная техника (2005).
8. Пискунов Н. С.: Дифференциальное и интегральное исчисления для втузов, т. 2: Учебное пособие для втузов. — 13-е изд.— М.: Наука, Главная редакция физико-математической литературы, (1985).
9. Сергеев А. Г. Крохин В. В. Метрология: Учебное пособие для вузов. М.: Логос (2000).
10. Тейлор Дж. Введение в теорию ошибок. Пер. с англ. – М.: Мир (1985).
11. Isolation Forest, https://scikit-learn.org/stable/modules/generated/sklearn.ensemble.IsolationForest.html (Дата обращения: 30.05.2020).
12. Local Outlier Factor, https://scikit-learn.org/stable/modules/generated/sklearn.neighbors.LocalOutlierFactor.html (Дата обращения: 30.05.2020).
13. Support Vector Machines, https://scikit-learn.org/stable/modules/svm.html (Дата обращения: 30.05.2020).

DIGITAL PLATFORMS
AND INFORMATION SYSTEMS

Formation of the Digital Platform for Precision Farming with Mathematical Modeling (Extended Abstract)

Victor Medennikov[0000-0002-4485-7132] and Alexander Raikov [1][0000-002-6726-9619]

[1] Federal Research Center "Informatics and Control" of the Russian Academy of Sciences, Vavilova 44-2, 119333, Moscow, Russia
[2] V.A. Trapeznikov Institute of Control Sciences of the Russian Academy of Sciences 65 Profsoyuznaya street, Moscow 117997, Moscow, Russia
dommed@mail.ru, alexander.n.raikov@gmail.com

Abstract. The paper addresses the issue of state and development trends of precision farming technologies (PFT) in the world. The main motive for the rapid development of PFT is the improvement of information and communication technologies (ICT), artificial intelligence methods, electron-optical equipment. It is shown that these technologies are currently evolving from the digitalization of individual operations to the digitalization of an interconnected set of operations in crop production and related industries. The approach makes PFT available for small and large farms. The paper analyzes the problems of PFT implementation such as following: the lack of a clear Ministry of Agriculture's strategy in this area, liquidation of the scientific organizations, the dominance of the "task-based" method of the development and implementation of digitalization systems, significant underutilization of traditional factors of increasing production efficiency in the industry, limited financial, labor, material and technical resources adapted for the implementation of digital technologies, poverty of most households. The paper discusses the scientific basis for the design of an optimal digital sub-platform for precision farming based on the mathematical and ontological modeling. An analysis is made of the constituent modules of a promising digital sub-platform of precision farming integrated into a unified geographic information space of the country. To demonstrate the possibility of forming a cloud service of the value chain, the concept of the single cloud Internet space for digital interaction of the logistic activities of agricultural production, processing and marketing of its products is presented.

Keywords: Digital Platform, Precision Farming, Mathematical Modeling, Geographic Information System.

In agriculture, the leading technology is precision farming technology (PFT). This approach utilizes cutting edge information technology to achieve the greatest improvement and efficiency in agricultural systems. It exploits modern information technology tools such as Global Positioning System (GPS), Geographical Information System (GIS), and Remote Sensing to improve farm management.

The essence of PFT is the integration of new agricultural technologies with the high-precision positioning based on remote sensing technologies (Earth remote sensing), as well as differentiated highly effective and environmentally friendly agricultural practices in the fields based on detailed information on the chemical and physical characteristics of each site. Digital PFT makes it possible to obtain the maximum number of products that have necessary requirements for quality, price, and safety.

Agriculture has to combine a huge amount of heterogeneous, multidimensional, diversified information with appropriate technologies for its processing. As a result, PFT has been evolved from the digitalization of individual operations to a complex of operations, moreover, not only in crop production but also with the integration of operations in related industries.

There are a lot of factors affecting both the industry itself and its digitalization. A scientifically integrated approach to the digital transformation of the industry based on mathematical modeling, taking into account financial, labor material, and technical resources, is required.

The main motivator for the rapid development of PFT in recent years was the improvement of ICT, electron-optical surveying equipment, the formation of global positioning systems that can simultaneously determine the coordinates of a significant number of objects with high accuracy anywhere in the world. Modern electron-optical equipment installed on various mobile (space, air, sea, transport, agricultural) and stationary devices has such a resolution that allows solving a significant range of problems in the field of agricultural production - from mapping the boundaries of individual land plots to the analysis of using intended to land and plant conditions over large areas.

Agriculture in terms of the production process looks and is usually described as a system of its interrelated elements in the form of resources used in production. To produce a certain type of product with given qualitative and quantitative characteristics, strict proportions between the elements (resources) of the system must be observed, due to the general and specific requirements of the production technologies of the planned products.

In the EU, almost all countries are beginning to use PFT, while Germany is the leader in PFT. Significant experiments are planned with PFT in Asia: China and India. In Germany, the Atfarm digital service is currently being tested. Atfarm is a cloud-based service in the field of PFT developed by the Norwegian manufacturer Yara for dosed plant nutrition based on plant health data collected using Yara N-Sensor equipment. This service helps farmers to apply nitrogen fertilizers using satellite remote sensing data at each specific site.

Precision farming is defined as a system consisting of interconnected subsystems: differentiated production technologies in crop production, a software and hardware complex for high-precision positioning of technological operations of the production process, a set of technical and agro-chemical means meeting the quality and quantity requirements. Therefore, the use of PFT should occur using an integrated approach. However, today in Russia's agriculture, a task-oriented approach to the design and development of information systems prevails, when they either order independently or purchase individual software systems that are ready from various manufacturers and

are neither connected ontologically, nor functionally, nor informational. This is also true with regard to the PFT. So, the first experiments of their use in agrarian production of Russia show their non-integrated, fragmentary use of individual technologies with filling heterogeneous information that differs in the structure during sending from one farm to another.

Following the trends of developed countries, the number of companies offering various individual solutions in the field of precision farming based on GIS technologies is growing in our country. Since these technologies are implemented in fragmented form, GIS is also unrelated in nature, capturing individual aspects of the processes. Moreover, the logical database structures are incomplete subsets of the ideal unified conceptual informational scheme of plant growth, which poses a threat to the future integration of agricultural information resources into a promising digital platform for the use of GIS. In this situation, following a task-oriented approach, the potential number of information systems in crop production that may appear in the absence of their integration may be assessed.

We think that about 150 topical tasks are to be solved in crop production for 20 crops, about 20 different technological operations are performed with each crop, and the number of regions is 80. Then, we can potentially get 4,800,000 information systems. From an analysis of the calculation results on the basis of a mathematical model of scenarios of possible options for informatization of the agricultural industry, it becomes clear that with this approach without abandoning task-oriented technologies for designing digital systems for standard and automated design with state support, the maximum possible level of the digital transformation of the industry will not exceed 17%. In addition, there are a significant number of other problems in the implementation of PFT technologies.

The most advanced approach to the design of digital systems determines the organization of life cycles and the life history of the "enterprise". An enterprise here is understood in a broad sense, for example, an enterprise may be a system for managing the industry, the economic sector, or the state. So, the use of this approach has found its application in the federal authorities of the United States, the construction of the electronic government of India.

So, an economic-mathematical model was developed for the formation of a digital platform for economic management, which allows us to calculate the optimal digital platform in agriculture. A digital platform is a business model for providing the possibility of an algorithmized exchange of information and values between a very large number of market participants by conducting transactions in a single information environment, leading to lower costs due to digital technologies and changes in the division of labor.

The model constructed by the authors made it possible to distinguish a number of digital sub-platforms, one of which is a cloud service for collecting and storing operational primary accounting information of all enterprises in the Unified Database of Primary Information (UDPI) in the following form: type and object of operation, place of implementation, subject of implementation, date and time interval conducting, means of production involved, the volume and type of resource consumed. The second is also a cloud service of a Unified Database of Technological Accounting of

all enterprises (UDTA). The ontological information model of crop production, based on them, is common for all agricultural enterprises in Russia with 240 functional management tasks with a single description of algorithms for most agricultural organizations (a standard for applications). Similar work was done for all sectors of agricultural production and 19 types of processing enterprises.

Thus, a digital platform is simulated, which is the integration of primary accounting information and technology databases in a single cloud environment. The TOGAF methodology is used. It is formed on the basis of a unified system for collecting, storing and analyzing primary accounting, technological, and statistical information, interfaced both with each other and with a unified system of classifiers, reference books, standards, representing registers of almost all material, intellectual and human resources of the agro-industrial complex.

The digital platform is acquiring special significance at the present time when technologies of remote sensing and GIS begin to be actively introduced in such a relatively young field of agricultural production as precision farming, which requires a combination of a large amount of data and technologies. The platform uses the possibility of Unified Geographically Distributed Information System of Earth Remote Sensing from space (UGDIS ERS).

UGDIS ERS is being created with the integration of all remote sensing information into a unified geographic information space of the country with an expiration date of 2025. The work is carried out in accordance with the plans of the development concept of the UGDIS ERS. Of course, it would be desirable to create the same center for decrypting images both from drones that are gaining popularity, as well as from stationary remote sensing masts, which would lead to a decrease in the cost of introducing PFT and increasing the efficiency of using these devices.

Remote sensing information after decryption in the created centers should be collected in a cloud GIS (CGIS), which also collects information on technological and primary accounting, data on all material, intellectual and human resources of the industry. An example of this approach is the existing in the EU Unified Administrative Management System (UAMS), which receives and stores information about lands and their users. Further, information obtained from sources other than those indicated above, as well as from gadgets, ground-based sensors, and sensors installed on agricultural machinery, is collected in a cloud GIS, while part of it is transmitted directly to communication equipment back to the equipment. Thus, the CGIS will collect all data on all technological and accounting operations performed at each site by all employees throughout the year. It will be possible to track all movements of products, materials and any equipment.

The digital platform based on PFT, GIS, and remote sensing technologies will create the basis of the operational management system. It will be a tool for economic analysis based on mathematical modeling, artificial intelligence, internet of thing, predictive logic, and big data in various sections at any level of management from the site to the federal center.

The consistent implementation of the promising PFT digital sub-platform will create the conditions for turning it into a set of scientifically-based infrastructure technologies for the entire agricultural sector.

Open Science Portal based on Knowledge Graph (Extended Abstract)

Vasily Bunakov[1][0000-0003-3467-5690]

[1] STFC UKRI, Harwell Campus, Didcot OX11 0QX, United Kingdom

The FREYA project [1] is developing the infrastructure for persistent identifiers (PIDs) and recommendations for their effective use as part of the European and global environment for Open Science. PIDs landscape is rich nowadays and includes identifiers for digital objects such as research publications or datasets, as well as for real-world entities such as people or organizations. The grand vision of FREYA is the "PID Graph" that creates relationships across objects and entities with PIDs and provides a basis for new services within research disciplines and across them.

The FREYA partners develop pilot applications that exploit parts of the PID Graph relevant to their respective research disciplines, also these applications in turn provide contributions to the PID Graph by opening up the information assets within the organizations using a common set of recommendations and where reasonable common technological solutions, too.

This work outlines a particular pilot application by STFC UKRI [2] in support of the Open Science agenda. STFC is a funder of science and of the post-graduate education in the UK, and a funder for the research conducted by the UK scientists on large-scale scientific instruments abroad. STFC is also a research organization that operates or co-owns large-scale scientific instruments (facilities) and high-performance computing across a few UK locations, granting access to them to the UK and overseas visitor scientists who conduct their own experiments.

All the mentioned streams of STFC funding and funding-in-kind (facility time and computation time awarded) result in research artefacts such as journal papers and preprints, PhD theses, data and software. Tracking down these artefacts back to the instruments, organizations and people involved is important for the evaluation of STFC role in a number of research fields such as biomedicine, chemistry, materials science, engineering, particle physics, astronomy, also of its role in higher education.

Research artefacts that have been produced with the support of STFC funding or funding-in-kind are reflected in records of science, e.g. bibliographic information for journal papers, or records of data deposited in certain reference databases. These records of science and the artefacts behind them are handled by a variety of information systems within STFC, also some well-curated STFC-related records of science are managed by external providers as parts of their larger collections (examples being crystallography or biomedical databases or national services for PhD theses).

To fully account for STFC funding and funding-in-kind, there is a need to systematically collect and manage these records of science, including the discovery of connections across them. Apart from this objective of having a better accountability for public spending on science, there is an important aspect of knowledge preservation and knowledge discovery in the spirit of Open Science that encourages and supports

reuse of research outcomes beyond the point of their origin. Open Science contributes to new research by other research organizations and individuals, raises the public awareness of science and its applications, also the organization itself can benefit from the more explicit and context-rich representation of its records of science with a single point of entry for their discovery.

Accountability and Open Science aspects are interrelated and can be supported by the same research information infrastructure that STFC are now building, with the first prototype of such infrastructure receiving support of FREYA project and having the focus on aspects that are specifically relevant to FREYA scope, i.e. persistent identifiers as a means of knowledge discovery and integration.

This new research information infrastructure is provisionally coined with the name of STFC Open Science Portal. This is going to be a publically available resource for the discovery of records of science that have been produced with the support of STFC funding or other flavours of sponsorship such as facility time awarded to visitor scientists. The records of science include information about research outcomes in any form (publications, data, etc.), records of STFC funding or other flavours of sponsorship, organizational context of STFC-supported research, as well as research attribution to particular large-scale instruments (facilities and their beamlines).

The Portal ingests metadata from a few sources, leaving data, full-text publications and other research artefacts in their current respective locations, and represents the integrated metadata as the knowledge graph. The metadata is subject to a moderate level of harmonization when integrated, yet there is no intention to support higher levels of the metadata harmonization or unification. Records of science in the external quality sources are not ingested in the Portal but linked from it through the use of persistent identifiers or in some cases by other record matching techniques.

The sources where the Portal ingests metadata from:

- STFC publications repository [4]
- STFC data repository [5]
- Diamond Light Source bibliographic database [6]
- DataCite (records there that have been produced by STFC) [7]
- Unpaywall [8] to discover Open Access versions of publications
- (under consideration) Gateway to Research [9]
- (under consideration) Crossref COCI and other sources of citations [10]

The sources that the Portal links to:

- The Cambridge Structural Database [11]
- The British Library EThOS service [12]
- Europe PMC [13]
- Protein Data Bank in Europe [14]
- (under consideration) Zenodo [15]

For managing the integrated metadata, a community edition of the neo4j graph database [16] is used that is currently hosted on the developers' computers and in the

STFC Intranet. The current size of the database is within tens of thousands of nodes and tens of thousands of relationships; this has a potential to grow to hundreds of thousands or a few millions of nodes, and hundreds of thousands or a few millions of relationships. The inflation of the database beyond hundreds of thousands or low millions of nodes and relationships is not expected at the moment.

For the records ingestion, OAI-PMH endpoints, bespoke APIs or bulk export features of the aforementioned sources have been used, which resulted in tabular or XML files. These have been further processed using XSLT transformations and Unix shell scripts, then uploaded in the graph database using standard neo4j tools.

Relationships across records are produced using elements of them associated with persistent identifiers where possible, otherwise fuzzy matching techniques are used, such as measuring distance between corresponding metadata elements, as was in the case of doctoral theses records [17]. The enrichment of scholar communication with persistent identifiers is a gradual process, and the broader PIDs proliferation should allow more efficient records matching in the future, saving the effort of building research knowledge graphs. This is a good example when best practices (of persistent identifiers minting and use) can augment and in certain cases replace technology (of fuzzy records matching).

Exploration and visualizations of the resulted graph and its parts (subgraphs) are made using queries in Cypher language [19] and standard Web components of neo4j. Further experiments with visualization are going to involve Apache ECharts [20] and JavaScript Cytoscape [21].

The vision of the Portal is for it to become a single entry point for searching across all records of science that could be publications, dataset descriptions, grants etc. The full-text search across all these records of science are supported by cross-records indexes created in the graph database and powered by Apache Lucene [18].

The development of the Open Science Portal has been focused so far on the data sources integration in the back-end graph database, with illustrative visualizations to see what usage scenarios are principally possible. The ultimate goal is to make the Portal a fully functional Web application with features of a full-text search across a variety of entities (publications, dataset and software descriptions, grants information) and with contextual visualization of metadata (subgraphs). The target audience for the graphical user interface are going to be STFC staff, visitor scientists, other funders and policy makers.

Apart from the graphical user interface, an API based on the GraphQL technology [24] is going to be implemented. Implementing a simple GraphQL endpoint to a graph database is straightforward and can be realized with standard plugins. A more sophisticated approach may be required though owing to the diverse nature of the metadata sources integrated; this might be based on measuring the popularity of certain attributes of the graph database nodes and relations [25]. The open API should allow third party developers to build their own applications around the STFC records of science.

The work is supported by FREYA project funded by the European Commission under the Horizon 2020 programme (Grant Agreement number 777523). The views expressed are those of the author and not necessarily of the project or the funder.

References

1. FREYA project, https://www.project-freya.eu/, last accessed 2020/09/08.
2. Science and Technology Facilities Council, https://stfc.ukri.org/, last accessed 2020/05/31.
3. UK Research and Innovation, https://www.ukri.org/, last accessed 2020/05/31.
4. ePubs: STFC publications repository, https://epubs.stfc.ac.uk/, last accessed 2020/05/31.
5. eData: STFC "Long Tail" data repository, https://edata.stfc.ac.uk/, last accessed 2020/05/31.
6. Diamond Light Source bibliographic database, https://publications.diamond.ac.uk/pubman/searchpublicationsquick, last accessed 2020/05/31.
7. DataCite search, https://search.datacite.org/, last accessed 2020/04/29.
8. Unpaywall: An open library of scholarly articles, https://unpaywall.org/, last accessed 2020/04/29.
9. Gateway to Research: UKRI gateway to publicly funded research and innovation https://gtr.ukri.org/, last accessed 2020/04/29.
10. COCI, the OpenCitations Index of Crossref open DOI-to-DOI citations, http://opencitations.net/index/coci, last accessed 2020/04/29.
11. The Cambridge Structural Database, https://www.ccdc.cam.ac.uk/solutions/csd-system/components/csd/, last accessed 2020/04/29.
12. The British Library EThOS service, https://ethos.bl.uk/, last accessed 2020/04/29.
13. Europe PMC portal, https://europepmc.org/, last accessed 2020/04/29.
14. Protein Data Bank in Europe, https://www.ebi.ac.uk/pdbe/, last accessed 2020/04/29.
15. Zenodo repository, https://zenodo.org/, last accessed 2020/04/29.
16. neo4j graph database, https://neo4j.com/, last accessed 2020/04/29.
17. Bunakov, V.; Madden, F. Integration of a National E-Theses Online Service with Institutional Repositories. Publications 8(2), 20 (2020). doi: 10.3390/publications8020020
18. Apache Lucene, https://lucene.apache.org/, last accessed 2020/04/29.
19. Cypher Query Language, https://neo4j.com/developer/cypher-query-language/, last accessed 2020/04/29.
20. Apache ECharts data visualization framework, https://echarts.apache.org/, last accessed 2020/09/08.
21. Cytoscape JavaScript library, https://js.cytoscape.org/, last accessed 2020/04/29.
22. Inelastic Neutron Scattering Database, https://www.isis.stfc.ac.uk/Pages/INS-database.aspx, last accessed 2020/04/29.
23. Plotly JavaScript Open Source Graphing Library, https://plotly.com/javascript/, last accessed 2020/04/29.
24. GraphQL query language, https://graphql.org/, last accessed 2020/04/29.
25. Bunakov, V. Metadata Integration with Labeled-Property Graphs. In: Garoufallou, E., Fallucchi, F., William De Luca, E. (eds.) Metadata and Semantic Research. Communications in Computer and Information Science, vol. 1057, pp. 441-448. Springer International Publishing, Cham (2019).

JOIN² Software Platform for the JINR Open Access Institutional Repository (Extended Abstract)

I. Filozova[1,2,3] [0000-0003-3441-7093], T. Zaikina[1] [0000-0003-0805-7995], G. Shestakova[1] [0000-0002-9826-8536], R. Semenov[1,3] [0000-0002-3203-5772], M.Köhler[4] [0000-0003-0617-3319], A.Wagner[4] [0000-0001-9846-5516] L. Baracci[5] [0000-0001-8433-948X]
on behalf of the JOIN² project

[1] Joint Institute for Nuclear Research, 6 Joliot-Curie St, 141980 Dubna, Moscow Region, Russia
[2] Dubna State University, Universitetskaya 19, 141982 Dubna, Moscow region, Russia
[3] Plekhanov Russian University of Economics, Stremyanny lane 36, 117997 Moscow, Russia,
[4] Deutsches Elektronen-Synchrotron DESY, Notkestraße 85, D-22607 Hamburg, Germany
[5] Deutsches Zentrum für Neurodegenerative Erkrankungen e. V. (DZNE), Venusberg-Campus 1, Gebäude 99, 53127 Bonn, Germany
fia@jinr.ru

Abstract. In recent years, Open Scientific Infrastructures have become an important tool for providing researchers and society with scientific information. Research institutes and universities worldwide actively plan and implement archives of their scientific output. Likewise, the JINR Document Server (JDS — jds.jinr.ru) stores JINR information resources and provides effective access to them. JDS contains numerous materials that reflect and facilitate research activities. Technically, JDS is based on the Invenio software platform developed by CERN.

To further improve the services, JDS is now adapting JOIN² workflows. JOIN², also based on Invenio, allows users, authors, librarians, managers, etc. to view the results of scientific work in a useful, friendly form and provides rich functionalities in the simplest way. The JOIN² workflow covers several verification layers of user data to minimize errors and thus provides checked and reliable information to end users. JDS features records with media files (video lectures, seminars, tutorials), and data import using DOI, ISBN, IDs from arXiv, WoS, Medline, PubMed, INSPIRE. Private collections with working group identification enable the integration into the research workflow. A common collection of Authority Records, i.e. grants, experiments, institutions, institutes, people, periodicals and their link with bibliographic records, establishes a high level of consistency and data quality.

Keywords: Open Access (OA), Institutional Repository, CRIS & OAR, JOIN², Invenio, Authority Records, JINR Document Server (JDS)

1 Introduction

The JINR Document Server (JDS — jds.jinr.ru) [1] is an information system representing an Open Access institutional repository (OAR) of articles, preprints and other materials that reflect and facilitate research activities at JINR. JDS is based on the Invenio software platform developed at CERN (formerly known as CDS ware and CDS Invenio). Invenio represents an all-inclusive application framework, which allows running a fully functional digital library server [2].

We consider JDS as a CRIS & OAR component of the JINR CIS. It reflects the results of the intellectual activity and must provide a record of all R&D activities, as well as retain the corresponding output. With the addition of the JOIN² (Just anOther INvenio INstance) [3] layer, it covers projects (funding), people (expertise), the organizational structure (groups), events, facilities and equipment (machines and experiments). As a repository, it provides a set of services to manage and distribute digital resources.

2 Implementation of the JDS prototype based on the JOIN² system

2.1 JOIN² project and prerequisites for migrating to JOIN²

JOIN² is a shared repository infrastructure that brings together eight research institutes for the development of a full-fledged scholarly publication database and repository based on the Invenio v1.1 open-source framework for large-scale digital repositories. The JOIN² members have consolidated their successful development workflow and collaboration and created a product that can meet the needs of a heterogeneous group of research centers [4]. JOIN² partners are the libraries of Deutsches Elektronen-Synchrotron DESY (Hamburg/Zeuthen), Deutsches Krebsforschungszentrum DKFZ (Heidelberg), Deutsches Zentrum für Neurodegenerative Erkrankungen DZNE (Bonn), the Joint Institute for Nuclear Research JINR (Dubna), Forschungszentrum Jülich (Jülich), GSI Helmholtzzentrum für Schwerionenforschung (Darmstadt), Maier-Leibnitz-Zentrum (Garching), Museum Zitadelle (Jülich) and Rheinisch-Westfälische Technische Hochschule Aachen (RWTH Aachen University). JOIN² enforces a well-defined publication workflow that is shared by all the project instances.

After some initial trials to flesh out the functionality of JOIN² for reuse in JDS, in 2017 JINR decided to become a partner [5].

The main challenges for adopting the JOIN² solution to JINR are related to:

- Multilingual authority records, e.g. to ensure seamless handling of authors' names in the Cyrillic and Latin script,
- adaptation of the default workflows to JINR issues,
- efforts to customize forms for easy data entry.

2.2 JDS-JOIN² Prototype

Keeping track of all (full coverage is the ideal case) publications of JINR employees is a significant challenge. There is an urgent need for a systematic approach to collect and preserve publications in a standardized way, to avoid multiple inputs of information and enable code reuse. The long-standing experience of the JOIN² partners is extremely useful and valuable for this activity.

Next, we will consider the main characteristics of the JDS-JOIN² prototype.

Software. The prototype is implemented using open-source technologies.

User interface. The Invenio platform provides a wide range of features for users.

Multimedia content. JDS embraces multimedia collections of video lectures for young scientists, posters, audio lectures, etc. Multimedia represents a crucial part of the content in the JINR institutional repository.

Submission of new records. The JOIN² system covers several levels of user data validation.

The submission of the publication comprises:

- Importing data from DOI, ISBN, arXiv, WoS, Medline, PubMed, INSPIRE;
- Enabling "exact matches" of the author based on a unique ID;
- Normalizing as much as possible through authority control;
- Several verification layers of user data provide reliable, consistent information to end users.

Authorities and Private Collections. Since it is planned to integrate JDS into the JINR corporate information system as a CRIS & OAR unit, it is essential to provide efficient and consistent reuse of data of employees, projects, grants, etc. Authority Records handle this task.

JDS Authority Records are:

- Grants,
- Experiments,
- Institutions and local institutes,
- People,
- Periodicals.

Some noteworthy features of JDS Authority Records are:

- Periodicals, Experiments, Grants and Institutions records form a unified base, which is shared between all the JOIN² partners,
- link with bibliographic records,
- access to Private Group Collections: only members of the contributing group are allowed to work with restricted materials.

*Search Generator and Advanced Search.*Simplifies the handling of authority-based queries and allows the user to experiment with different kinds of queries, including default exact word matching, phrase searching and regular expression matching, as well as filtering on specific indexes such as “fulltext”, “issn”, “experiment”.

Statistics and Reporting. JOIN² encompasses a module for statistics and reporting with export to PDF. The statistics tool enables the creation of various types of reports that reflect and facilitate research activities of the institute. It is noteworthy that it can be used by scientists to handle publication lists or citations in their work, as well as by staff for reporting.

Export of publications. Users can export a list of their publications into their website or to various bibliographic formats. In addition, JOIN² provides the ability to export a list of publications within an experiment or by a research group as a URL and to include this list on a website.

3 Conclusion

JOIN² aims at making information and knowledge accessible and visible through a unified information hub-like solution, which all partners use. A distinctive feature and practical value of the project is that partner institutions join forces to enrich the functionality of the software platform and share Authority Records to enhance data quality, reduce human errors and eliminate redundant work. As a result, unique experience and competence are accumulated; an interactive enrichment of local teams takes place.

JOIN² covers all the aspects of JINR needs for an institutional repository. There is no doubt that the JDS-JOIN² prototype of an institutional repository, developed as an extension of JOIN², has significant advantages over the previous solution of JDS, which was based on Invenio without JOIN² enrichments. Several user groups have already provided a positive feedback on the JDS-JOIN² collaboration. The prototype is an evolutionary process, and the production system will be finalized in the near future. We believe that the JDS-JOIN² implementation will satisfy the information needs of target users.

4 References

1. JINR Document Server, http://jds.jinr.ru, last accessed 2020/09/04.
2. The official web-site Invenio project, https://invenio-software.org/, last accessed 2020/09/04.
3. JOIN² Project, https://join2.de, last accessed 2020/09/04.
4. Baracchi L., Wagner A. JOIN² a Publication Database and Repository Based on Invenio. In: Korenkov V., Strizh T., Nechaevskiy A., Zaikina T. (eds.) The 27th Symposium on Nuclear Electronics and Computing, NEC 2019, CEUR-WS.ORG, vol: 2507, pp. 51-67. RWTH Aachen (2019), http://ceur-ws.org/Vol-2507/51-57-paper-8.pdf.
5. Wagner, A., Invenio @ JOIN², 4th Invenio User Group Workshop, IUGW2017, Garching, Heinz Maier-Leibnitz Zentrum, Germany, 21 - 24 Mar 2017, doi:10.3204/PUBDB-2017-01356.

Innovative Approach to Updating the Digital Platform Ecosystem (Extended Abstract)

Alexander Zatsarinnyy [1] and Alexander P. Shabanov [1]

[1] Federal Research Center "Computer Science and Control" of the Russian Academy of Sciences, Vavilova st. 44-2, 119333 Moscow, Russia
apshabanov@mail.ru

Abstract. A new method is proposed for generating project scenarios in the configuration management database to ensure interoperability with the new organizational system. This article is intended for researchers and developers. The study is part-supported by the RFBR, projects 18-29-03091, 18-29-03100.

Keywords: organizational system, ecosystem, digital platform, interoperability, database, project scenario.

1 The production of a task

The research relates to the solution of the scientific problem of economic development in the direction of "new production technologies". This study is preceded by the authors' work on solving problems related to decision support, information interaction, and critical technology and resource management, for example, [1-4]. The objects of control in them, as in this study, are configuration management databases containing records of configuration elements of the systems under study. The purpose of this research is to develop a new method for creating project scenarios in the configuration management database to ensure information interaction with a new organizational system that is being introduced into the digital platform ecosystem

2 Methods and models for developing project scenarios

The main methodological approach used in the study of the problem is the concept of it service management.

As a result of the analysis of information on the practice of applying this concept in organizational systems of various sectors of the economy, a logical causal sequence of properties of the configuration management database is revealed, which allows us to present it as an electronic model of information infrastructures in existing organizational systems. An elementary unit in such models is the configuration record. The configuration management database is essentially a universal component capable of methodically ensuring the functioning of organizational systems in digital circuit

Board ecosystems, regardless of their number and activities of organizational systems. The possibility of using data processing and transformation technologies - Information Centric Network, and technologies for data transmission, reception and transformation - Intent-Based Networking to automate processes reproduced in the digital platform ecosystem in the direction of developing new production technologies was determined.

As a result of the analysis of well-known research and development works (R & d), startups and intellectual property objects, which precede technical and working projects in the direction of "new production technologies", a range of technological directions (robotic systems, virtual and augmented reality, artificial intelligence, big data), which covers all key business areas. Thus, the information character of supporting business processes and information interaction processes, which is inherent in the methods (table 1), allowed us to apply it service management processes in accordance with the concept of IT Service Management to prove their industrial feasibility.

Table 1. Examples of patents to support organizational processes

Title	Purpose
1. Method of supporting operation of organizational system. (https://new.fips.ru/Archive/PAT/2014FULL/2014.11.10/DOC/RUNWC2/000/000/002/532/723/DOCUMENT.PDF)	This improves the efficiency of information support by automatically executing scenarios for evaluating the performance of configuration elements and automatically managing them based on the completed evaluation.
2. Method of transmission of control commands. (https://www1.fips.ru/ofpstorage/Doc/IZPM/RUNWC1/000/000/002/631/147/И3-02631147-00001/document.pdf).	Interoperability is achieved by converting data about the control command team to data about its function and converting it back to the no in other codes.
3. Method of information transmission. (https://www1.fips.ru/ofpstorage/Doc/IZPM/RUNWC1/000/000/002/618/366/И3-02618366-00001/document.pdf).	Interoperability is provided on the transmitting side by converting data about the information to data about its destination, on the receiving side by reverse conversion, but in other codes.

Based on the results of the analysis, it can be noted that the data sources in the processes of processing, transformation... are configuration management databases, and processes for ensuring information interaction can be classified as processes that use configuration records - project scenarios about configuration elements in management systems, information systems, and the digital platform ecosystem.

3 Formal interoperability model in the configuration management database

Based on the results of the analysis and taking into account the assumption that the configuration record is presented as a project scenario, a formal model was developed to describe the segment of information interaction in the configuration management database of the digital platform ecosystem. The model has the following levels.

The first level is the system level, where $S = \{S_1, S_2 \ldots S_N\}$ is a set of configuration entries about configuration elements that represent organizational systems, and N is the number of organizational systems in the ecosystem.

With each of the subsets $S_1, S_2 \ldots S_N$ refers to the configuration elements of the internal information infrastructure and the configuration elements of the external environment - data transmission networks, engineering structures, etc. When placing configuration records about these elements in the configuration management database, they are converted to exclude duplicate records. These are usually entries about elements of the external environment.

The second level is individual, in which $V = \{V_1, V_2 \ldots V_M\}$ is a set of configuration records about M classes of information interaction objects. For example, there are differences in component indexing systems (V_1), component program codes (V_2), data encoding systems (V_3), and information encryption systems (V_4).

The third level is an object level, where $V_i = \{V_{i1}, V_{i2} \ldots V_{iMi}\}$ is a subset of configuration records about configuration elements that represent objects for providing information interaction for class V_i, where Mi is the number of objects in this class.

The third-level set $U_{DPE} = \{U_{OS}, U_{DP}\}$ is a non-repeating configuration record of U_{OS} configuration elements used for reproducing organizational system processes and U_{DP} elements used for reproducing platform processes.

This property of the model is crucial for operations to update the digital platform ecosystem by introducing a new S_{N+1} organizational system.

4 Method for generating project scenario data in the configuration management database

The method of generating data about the project scenario in the configuration management database is characterised by the following actions:

- accept and save data about the project scenarios of the ecosystem and the new organizational system;
- conduct a comparative analysis of this data;
- check the identity condition;
- if the condition is not met, form and save data about the new project scenario as data to be written to the database;
- when executed, link this data to records about the new organizational system.

The scientific novelty of the method is to provide automatic generation of records about configuration elements of a new business entity in the configuration management database in the digital platform ecosystem. The practical significance of the study is to reduce the time required to build competencies and resources of the digital platform ecosystem by introducing new organizational systems.

5 Conclusion

For the first time, the article sets the task of finding a methodological approach to updating the configuration management database by using information about interaction with a new business entity introduced into the digital platform ecosystem. To solve the problem, we use the methodology of IT Service Management and the best international practices of Information Centric Network, Intent-Based Networking. The analysis of innovative methods and models to support the development of new production technologies in the digital platform ecosystem is performed. A multi-level formal model for describing the segment of information interaction in the configuration management database has been developed and a new method for generating data about the project scenario in it has been proposed.

References

1. Zatsarinnyy, A.A., Shabanov, A.P.: Situational Centers: information – processes – organization. Telecommunications, 2011, no 1, pp. 84-87.
2. Zatsarinnyy, A.A., Kozlov, S.V., Shabanov, A.P.: Interoperability of Organizational Systems in Addressing Common Challenges. Management Issues, 2017, no 6, pp. 43-49.
3. Zatsarinnyy, A.A., Shabanov, A.P.: Information Support for the Activities of the Critical Technologies in Control Systems Based on Situational Centers. Systems of Control, Communication and Security, 2015, no 4, pp. 98-113.
4. Zatsarinnyy, A.A., Shabanov, A.P.: Models and methods of cognitive management of digital platform resources. Control Systems, Communications and Security, 2019, no. 1, pp. 100-122.

Система экологического мониторинга в рекреационных зонах на основе больших данных (Расширенные тезисы)

А. Н. Волков[1], А. С. Копырин[1], Н. В. Кондратьева[1], С. С. Валеев[1]

[1] Сочинский государственный университет, Сочи, Россия
vss2000@mail.ru

Аннотация. Рассматривается архитектура системы экологического мониторинга рекреационной зоны, в которой используются технологии больших данных для обработки информации, полученной с помощью дистанционного зондирования земли, аэрофотосъемки, съемки с беспилотных летательных аппаратов, внутреннего мониторинга состояния объектов рекреационной зоны, а также статистические данные.

Ключевые слова: Экологический мониторинг, Большие данные, Многоуровневая система сбора информации, Базы данных, Агрегирование данных.

Как известно, основной целью рекреационной зоны является восстановление здоровья населения и обеспечение сохранности трудового капитала любого государства. Оценка загрязнения рекреационной зоны с учетом масштабов загрязнения окружающей среды позволяет контролировать экологическую ситуацию в рекреационной зоне в пределах региона.

Оценка состояния сложных систем основана на иерархическом наборе моделей различной степени детализации [1]. При построении системы мониторинга экологического состояния рекреационной зоны также используются модели различного уровня, основанные на анализе имеющейся информации.

Тепловые выбросы в окружающую среду и выбросы вредных веществ мегаполисов приводят к значительному ухудшению экологического состояния курортной зоны, находящейся достаточно далеко от мест формирования вредных выбросов. Эти обстоятельства влияют на уровень предоставляемых рекреационных услуг, а также экологическое состояние курортной зоны. Построение прогноза экологического состояния приморских рекреационных зон, управляемое снижение вредных выбросов различной природы является чрезвычайно актуальной научной и практической задачей [2, 3].

Определение площади зон повышенного содержания вредных веществ позволяет качественно оценить уровень загрязнения зоны на основе экспертных оценок и аналитики больших данных. Суммарная площадь зон с повышенным уровнем вредных веществ в разное время суток и время года значительно отличается. Это обусловлено рядом причин, среди которых климатические условия

рассматриваемого рекреационного региона и траектория движения воздушных масс, переносящих на большие расстояния вредные вещества [4-7].

Из этого можно сделать вывод, что для повышения качества прогноза необходимо учитывать динамику экологического состояния региона с учетом всех доступных данных, формируемых метеослужбами и данных, полученных на основе дистанционного зондирования земли в различных спектральных диапазонах.

Рассматривается задача интеграции систем сбора и обработки большого объема данных, решение которой позволяет учитывать множество различных факторов, влияющих на динамику экологического состояния [8 - 11].

Для эффективного решения рассматриваемой задачи предлагается разработка многоуровневой системы сбора информации об экологических характеристиках рекреационной зоны [12 - 14]. В данной системе сбора и обработки данных интегрируются массивы данных, получаемые на основе обработки информации спутниковых снимков, данные распределенного мониторинга теплового состояния объектов инфраструктур зоны, а также данные об уровне концентрации вредных веществ в рекреационной зоне в различное время суток и время года.

Использование технологии больших данных позволяет решить задачу интеграции большого объема данных о тепловых потерях, актуальных данных об уровне загрязнения воздуха, оцениваемого на основе анализа спутниковых изображений, результатов моделирования экологической обстановки, информации о степени загрузки транспортных артерий региона, и данных о розе ветров и текущих климатических условиях.

Рассматриваются особенности архитектуры трехуровневой системы сбора данных мониторинга движения воздушных масс в регионе, а также системы сбора данных об уровне загрязнения воздуха. Система агрегирования данных включает в себя подсистемы обработки данных, получаемых на основе анализа изображений и данных метеонаблюдений.

Таким образом, рассматриваемая проблема получения достоверного прогноза связана в первую очередь с эффективной организацией сбора информации об экологическом состоянии объектов курортной зоны. Сложность решения рассматриваемой задачи обусловлена необходимостью сбора, обработки и хранения разнородной информации, а также необходимостью решения задачи, связанной с разработкой прогноза экологического состояния региона в условиях влияния различных факторов неопределенности.

Библиографический список

1. Thalheim, B.: Models for Communication, Understanding, Search, and Analysis. In: Proceedings of the XXI International Conference on Data Analytics and Management in Data Intensive Domains (DAMDID/RCDL 2019), pp. 3-18 (2019)
2. Overall heat transfer loss from buildings - transmission, ventilation and infiltration, https://www.engineeringtoolbox.com/heat-loss-buildings-d_113.html
3. Theodore, L., Behan, K.: Introduction to Optimization for Chemical and Environmental Engineers, 1st ed. CRC Press (2018)

4. Top six places for energy losses in commercial buildings. https://www.ecdonline.com.au/content/test-measurement/article/top-six-places-for-energy-losses-in-commercial-buildings-259097222
5. Building Envelope: How to Avoid Energy Loss. https://www.facilitiesnet.com/energyefficiency/article/Building-Envelope-How-to-Avoid-Energy-Loss--9428
6. Aerial Infrared Inspection. http://nationwidedrones.co.uk/drone-infrared-inspection
7. Making Your Facilities A Safer Place. http://preciseir.com/
8. Leskovec, J., Rajaraman, A., Ullman, J.D.: Mining of massive datasets. Cambridge University Press (2014)
9. Bohlouli, M., Schulz, F., Angelis, L., Pahor, D., Brandic, I., Atlan, D., Tate, R.: Towards an integrated platform for Big Data analysis. In: Integration of Practice-Oriented Knowledge Technology: Trends and Prospectives. Springer, Berlin, pp. 47–56 (2013)
10. Lenzerini, M.: Data integration: a theoretical perspective. In: Proc. of the 21st ACM SIGACT SIGMOD SIGART Symp. on Principles of Database Systems (PODS 2002), pp. 233–246 (2002)
11. Sazontev, V.: Methods for Big Data Integration in Distributed Computation Environments. In: Proceedings of XX International Conference “Data Analytics and Management in Data Intensive Domains” (DAMDID/RCDL’2018), pp. 239-244 (2018)
12. Kondratyeva, N.V., Valeev, S.S.: Simulation of the life cycle of a complex technical object within the concept of Big Data. In: CEUR Proceedings of 3rd Russian Conference Mathematical Modeling and Information Technologies, pp. 216-223 (2016)
13. Волков А.Н., Копырин А.С., Кондратьева Н.В., Валеев С.С. Информационная система экологического мониторинга в рекреационных зонах. Научный журнал. Инженерные системы и сооружения. 2020. Т. 1. № 1 (38). С. 20-24.
14. Волков А.Н., Копырин А.С., Кондратьева Н.В., Валеев С.С. Система сбора информации об энергетических потерях в рекреационных зонах. В книге: Управление развитием крупномасштабных систем MLSD'2019. Материалы двенадцатой международной конференции Научное электронное издание. Под общей ред. С.Н. Васильева, А.Д. Цвиркуна. 2019. С. 838-841.

New Approaches for Delivery of Data and Information Products to Consumers and External Systems in the Field of Hydrometeorology (Extended Abstract)

Evgenii D. Viazilov[1], Denis A. Melnikov[1] and Alexander S. Mikheev[1]

[1] RIHMI-WDC, 6, Koroleva St., 249035. Obninsk, Russia
vjaz@meteo.ru

Abstract. The delivery of data and information to consumers by subscription e-mail or by placing it on an ftp server for further upload to external information systems is currently one of the most sought functions of information systems. In the field of hydrometeorology, this function is using in individual organizations and systems, requires a more systemic approach to its development for information support to consumers. An analysis of the methods of delivery data to consumers is presenting. The features of data automatic delivery to external systems and consumers using the Unified state system of information on the condition in the World Ocean are considered. New for hydrometeorology approaches automatic delivery of information about the different threats, visualizing the state of the current situation indicators in the form of a dashboard, the data delivery to external information systems of the enterprises with automatic loading into the database are proposed. The full implementation of these approaches will make it possible to organize the more effective hydrometeorological support for consumers. Prospects are development of indicators for evaluating the delivery of data (number of consumers are on service; number of transferring, etc.).

Keywords: Threat Identification, Data Transferring, Disaster Indicators.

1 Introduction

A data dissemination system and of information products in the field of hydrometeorology are standardized at the international level in the framework of the global telecommunication system (GTS), which operates under the auspices of the World meteorological organization. All countries contribute to the GTS with their observations and information products, based on observed data. The amount of information circulating in the GTS is tens of GB per day. In each country develops its own system of hydrometeorological support of consumers which use observational data, as well as global and regional forecasts issued by leading centers of forecasting and transferred by GTS.

Delivery data and information to consumers is a process consisting in converting of data flow, information that affects the course of this process into a form that ensures prompt and error-free perception by the consumer and the direct issuance of information. This term are using in standard ISO 29481-1:2016 (Building information models: Information delivery manual. Part 1: Methodology and format). In hydrometeorology, the term "Delivery of data and information" is understood as:

- Transfer of hydrometeorological data to heads of enterprises, representatives of public organizations and population on a regular basis
- Prompt notification of government officials and the public about the emergence of threats in the form of natural disasters
- Interaction of the population and enterprises heads with interactive applications to obtain detailed information about the state of the hydrometeorological situation.

2 Subscriptions of Consumers on the Requested Data

Subscribing to data delivery is a most sought function of many systems. Subscriptions are usually using for the regular delivery of data / metadata files or links to geoservices. In order to increase the using data effectiveness, it is necessary to switch to automatic data delivery provided when threshold values of threats indicators are exceeding. For this is required:

- To develop a technology for by the automatic delivery of information to heads of enterprises, public authorities and the public, for assessment of the impact of threats, and the possible changes in climate in the various economic sector
- To develop tools of threats estimates for the levels of danger for enterprises, types of activity with links to more detailed consideration of situations (maps of forecast, observed parameters)
- To carry out automatic delivery of messages about threats and information on exceeding the threshold values of indicators of threats
- To develop a program-agent to automatically launch application on mobile Internet devices of heads and inform consumers in case of exit of values of parameters behind the threshold values
- To transfer to the consumers required current, and or prognostic, or climate information with parameters for the particular industrial enterprise not on the initiative of the consumer, and by automatic delivery on any Internet device.

For hydrometeorological support, including subscription, it is necessary to have coordinates of the point or region, the composition of parameters; the type of information, the type of object, type of activity, threshold values of indicators of threats, e-mail address or the number of mobile phone, to which it need transfer the data. This information should allow determining geographical area; category of data (observation, analysis, forecast, climate); data period (only last arrivals, last receipts + forecast, last receipts+ forecast + climate); for what the necessary data (replenishment of times series, tables with analytical and or prognostic data, decision-making, et al.); threshold values of indicators for various types of enterprises, types of activities, for every threats; information resources on the state of the environment, based on which threats determine.

For each indicator, algorithms and methods of preparation are developing. The organization scheme for hydrometeorological support of heads in case of threats is the increasing information of enterprise heads, quick familiarization with the current situation by:

- Automatic calculating of indicators based on of identifying various threats in the form of "traffic light" of various danger levels (green, yellow, orange, red)
- Selection and delivery of received storm alerts
- Delivery of information on detected threats, including providing of information on the impacts and recommendations for decision support
- Visualization of the hydrometeorological situation with the help of digit, maps, pictures, icons, defining generally recognized hydrometeorological instruments
- Detailed acquaintance with the hydrometeorological situation by maps, figures and tables
- Delivery of information into the information systems with automatic loading into database
- Receiving of information about possible damage and cost of preventive actions before disaster.

3 Applications for Determining the Level of Dangerous

This tool is designing to identify threats based on select of values in flow of observed and forecast data, which exceed thresholds. Identification are making based on information resources, for the selected object and threshold values of threats. The result of work this application is a constantly updated database with dangerous situations for each object.

For the functioning of the threat monitoring system, the use of integrated data obtained from existing data sources is necessary. The input information to determine impacts of threats on objects is the values of the measured parameters at the point (fixed hydrometeorological station, or of a regular grid).

The database creation of threshold values of threat indicators is performed according in the form of "traffic light", taking into account the type of enterprise. Threats are determined both at the level of unit of observations (excess of indicator recorded in one point), so and the region (the threat is registered in several points of observations). When this is using as warnings of the territorial offices of Roshydromet of threats, so and threats, which identified automatically based on observed and forecast data. In result, head is no need to be constantly at the computer in order to monitor the situation. Color and sound signaling helps to the timely to draw attention to the message received. Decreasing amount of data, provided to the head in the message window of the threat, makes it easier acquaintance with the situation. It is proposed to create such a system in work [2].

4 Tools of Delivery of Information for Threats

4.1 Application of the Selection and Transmission of Storm Warnings

For hydrometeorological support, it is necessary, first, to use the official "Storm warnings and alerts" of Roshydromet organizations. These warnings are compiled by observers on stations or synoptic in territorial offices of Roshydromet, transmitted via GTS. In first case of information about threats identified by observers at the hydrometeorological station and transmitted in code WAREP via the GTS. These messages can be using to identify objects that threats affect.

The second option is, when the synoptic in the territorial offices of Roshydromet and of research institutions, based on analysis of the current situation, weather maps and other materials, are forecast threats and pass it to interested enterprises.

These threat forecasts and warnings be using to identify objects that they are influencing. Identified messages for a specific object (district) are stored in the database and must transfer to consumers using SMS.

4.2 Application for Delivery Information about Identified Threats

Application for automatic delivery of information about the various threats is intending for visualization of information about threats based on the received SMS with the address of the software for link [1, 3]. The message includes the name of the object, the name of the indicator, its value, and dangerous level. The message contains a link to the application for more detailed information about the current hydrometeorological situation. In addition to impacts and recommendations, the consumer can be able to assess possible damage and calculate the cost of preventive actions. This application is installing on the smartphone of enterprise head and is activating when SMS is receiving. The application should conduct monitoring of alerts about threats in real time. The availability of information on mobile phone numbers registered at base stations located in the threat zone allows providing prompt alerts and instructions the owners of mobile phones.

4.3 Application in the Form of a Dashboard

In addition to SMS messages, the head should see the state of indicators of the current situation. To do this can use the application to form indicators as a dashboard. The application must in an interactive mode to show values of observed and prognostic parameters celebrated on icons meteorological instruments, with indication of the level of dangerous. Beside each, parameter value should be the graphic changes of anomalies and of trends. In such an application, information is displaying on the screen in a more compact form. One glance at this form will be enough to understand the current hydrometeorological situation. If necessary, the head can get detailed information on the ESIMO portal[1] for the district by application MeteoMonitor.

[1] http://esomo.ru - Unified State System of Information on the Condition in the World Ocean

4.4 Application MeteoMonitor

Application MeteoMonitor is software for more detailed acquaintance with the hydrometeorological situation at a point or in a spatial. The program interface of this application should provide the following forms of presentation of information: maps of the spread of threats in spatial; graphs of changes in threat indicators in form time series; tables of values of environmental parameters in a particular observation point or closest point on regular grid; indication of values, showing the status of indicators of threats, for individual objects and types of activities on them; messages about threats; warning about disasters via sound, color.

4.5 Application for the Delivery of Information to the IS of Enterprises with Automatic Loading into the Database

At the present time it is necessary to hydrometeorological data are using in the automated business processes of enterprises. Depending on the business process, data can be transferred for a point, region or trajectory.

If the enterprise is to a point, then the observed hydrometeorological information available from hydrometeorological station, which is located from the enterprise closer, and prognostic information - on the nearest point of grid. If the enterprise has distributed in spatial, then for each object selected by points of observations and results of interpolation at the grid point, how, and in the previous case. If the enterprise is a dynamic object (vessel), then observation points and grid points along the entire route of the vessel are selected for it.

5 Conclusion

New for hydrometeorology approaches automatic delivery of information about the different threats, visualizing the state of current situation indicators in the form of a dashboard, the data delivery to external information systems of the enterprises with automatic loading into the database are proposed. Full implementation about these approaches must permit to organize that is more effective hydrometeorological support of enterprises.

Acknowledgements

The work financially supported by the Ministry of Science and Higher Education of the Russian Federation, a unique project identifier RFMEFI61618X0103.

References

1. RD 52.27.881 Guide to the hydrometeorological support of marine activities. M., Hydrometeorological Center of Russia, 132 (2019).

2. The revised road map for the Decade of the Organization of the United Nations, dedicated to the science of the ocean in the interests of sustainable development. IOC / EC-LI /. Annex 3. Paris, 18 June 2018. Original: English. UNESCO Intergovernmental Oceanographic Commission. Fifty-one Session of the Executive Board of UNESCO, Paris, 3-6 July 2018, 66 (2018).
3. Viazilov, E.: Development of hydrometeorological services to support decisions of enterprise leaders: Examples from the Russian Federation. United Nations Office for Disaster Risk Reduction (UNISDR). The UN Global assessment report on disaster risk reduction 2019 (GAR19). Contributing Paper to GAR 2019 (Global Assessment Report on Disaster Risk Reduction), 26 (2019). https://www.preventionweb.net/publications/view/66441/.

DATA ANALYSIS IN MEDICINE

EMG and EEG pattern analysis for monitoring human cognitive activity during emotional stimulation* (Extended Abstract)

Konstantin Sidorov [1], Natalya Bodrina [2] and Natalya Filatova [3]

[1, 2, 3] Tver State Technical University, Lenina Ave. 25, Tver, Russia
bmisidorov@mail.ru, vavilovani@mail.ru, nfilatova99@mail.ru

Abstract. The paper describes the experiments on monitoring human cognitive activity with additional emotional stimulation and their results. The purpose of the research is to determine the characteristics of EMG and EEG signals that reflect an emotional state and cognitive activity dynamics. The experiments involved using a multi-channel bioengineering system. The channels for recording EEG signals (19 leads according to the 10-20 system), EMG signals (by the "*corrugator supercilia*" and "*zygomaticus major*" channels according to the Fridlund and Cacioppo methodology) and the protocol information channel were engaged. There is a description of an experimental scenario, which assumed that testees performed homogeneous calculating tasks. According to the experimental results, there were formed 1344 artifact-free EEG and EMG patterns with a duration of 4 seconds. During emotiogenic stimulation, an EMG signal by the corresponding channel intensifies and a power spectrum shifts to the low-frequency region. An emotional state interpreter based on a neural-like hierarchical structure was used to classify EMG patterns. The classification success was 93%. The authors have determined spectral characteristics and attractors of EEG patterns. The highlighted attractor features were: the averaged vector length for the *i*-th two-dimensional attractor projection; density of trajectories near its center. The most informative frequency range (theta rhythm) and leads (P3-A1, C3-A1, P4-A2, C4-A2) were selected. These features have revealed a decrease in testees' cognitive activity after 30-40 minutes of work. After negative emotional stimulation, there was an increase in absolute power in the theta rhythm, an increase in the average vector length for the *i*-th two-dimensional attractor projection, and a decrease in the trajectory density in four central cells. Tasks success indicators were improving. The revealed EEG signal features allow assessing the current cognitive activity of a person taking into account the influence of emotional stimulation.

Keywords: cognitive activity, emotional stimulation, bioengineering system, biomedical signal, EEG, EMG, attractor.

* The work has been done within the framework of the grant of the President of the Russian Federation for state support of young Russian PhD scientists (MK-1398.2020.9).

1 Introduction

Nowadays, much attention is paid to the relationship between human emotional state and the effectiveness of performing cognitive tasks of various types. There is a need in quantitative estimates of emotional arousal and cognitive congestion of the brain. For this purpose, the researchers use various biomedical signals. Signals of electrical activity of the brain (EEG) is now considered to be very informative [1-5]. However, there is no clear delineation between signals reflecting emotional state and cognitive activity [6]. For this reason, it is necessary to use combined multi-channel systems in studies. Facial electromyography (fEMG) is widely used to determine the emotional state of a person [7, 8]. To classify emotional states according to EMG, deep learning methods are used. The level of cognitive activity is assessed by the results of tasks performing. However, there are known results that describe the activation and interconnection of various structures of the human brain when solving various types of problems [9-11].

The purpose of our research is to highlight the features in EEG and EMG signals in order to form an attributive model of cognitive activity monitoring taking into account the effect of emotional stimulation.

2 The Experiment

A series of experiments was carried out using the "*EEG-Speech+*" system [12] with three channels: electroencephalograph (Encephalan 131-03, Russia), myograph (Neuro MVP-4, Russia) and a protocol channel. EEG recording was done according to the 10-20 system by 19 leads with a sampling frequency of 250 Hz. EMG was recorded in the "*corrugator supercilia"* and "*zygomaticus major"* channels; a ground electrode was placed on the center of the forehead. Recording had a sampling frequency of 1000 Hz.

The participants were 12 testees at the age of 20–30. During the experiment, a testee performed 700 computational tasks. The protocol had records of all testee's answers and the time spent on each group of 10 tasks. The stimulation included a video lasting 20 min. Each testee had two experiment sessions, with positive and negative stimulation. We selected 56 EEG and EMG patterns lasting 4 seconds each free of artifacts from each experiment session. The EMG records were filtered through a bandpass of 20-500 Hz and a notch filter to suppress a 50 Hz network component.

3 Analysis of EMG patterns

We used a sliding time window without overlapping lasting 4 seconds. The ratio of the root mean square (RMS_i) for time window i and the average RMS_{sr} fragment with open eyes, when a testee did not do any tasks, was the main sign of emotional changes. The second sign was the frequency change (median) for the i-th time window, which divides the power spectrum into two parts equal in total intensity.

With stimulation, the signal from the corresponding muscle was significantly amplified and its spectrum shifted toward low frequencies.

In [13], an interpreter of a human emotional state based on a neural-like hierarchical structure (NLHS) was proposed. Based on training sets analysis a several variants of NLHS was created and classification rules for emotion classes were retrieved. The obtained results demonstrated an 93% accuracy for classification of EMG patterns.

4 Analysis of EEG patterns

Firstly, we used the methods of spectral analysis. For all EEG fragments we determined *S(f)* power spectra – power spectral densities (PSD_s), using the Fourier transform and Welch's method. We used the Hamming window (width 512), the frequency range varied from 0 to 125 Hz. The formalized description of an arbitrary EEG pattern:

$$S(f)_{\text{EEG}} = \langle \{f_1, f_2, \ldots, f_r\}_l \rangle, \quad (1)$$

where $S(f)_{EEG}$ are *PSD* feature vectors; *f* is an EEG object number ($f = 1 \div 1344$); *l* is an EEG lead number ($l = 1 \div 19$); *r* is *PSD* feature number ($r = 1 \div 250$), *PSD* calculation step is 0.5 Hz. For each calculated feature vector (1), we determined the absolute power value (*AP*, $\mu V^2/Hz$) – the area under the corresponding *PSD* section by selected frequency ranges. Further, we used the attractor reconstruction procedure [14].

During the analysis of $AP(\Delta f)$ features, the most informative EEG leads were localized. They are P3-A1, C3-A1, P4-A2 and C4-A2. Particular attention was paid to the selection of useful frequency rhythms illustrating a change in human cognitive activity. The experiments showed that the testees had maximum changes in the theta rhythm.

Thus, to assess testee's cognitive activity by one EEG signal lead, we used 3 feature types: 1) Pr1 – averaged vector length for *i*-th first two-dimensional attractor projection; 2) Pr2 – trajectory density in 4 central cells of the *i*-th attractor projection; 3) Pr3 – absolute power at theta rhythm interval (4-8 Hz).

Monitoring of the features Pr1, Pr2, Pr3 revealed a decrease in cognitive activity at about the 35th–40th minute of the experiment. Theta rhythm after a negative emotional stimulation showed increasing Pr3. Also, we revealed such stimulation causes an increase of Pr1 and a decrease of Pr2. Finally, we can observe a decrease in the number of mistakes and an increase in the speed of completing tasks.

5 Conclusion

The research has established that the EMG signal makes it possible to quickly identify testees' emotional state. The identification accuracy is 93%.

After video stimulation, which caused weak negative emotions, we can observe an increase in absolute power values in the theta rhythm, as well as a decrease in the

number of mistakes and an increase in the speed of completing tasks. The research has revealed an increase in the averaged vector length for the *i*-th two-dimensional attractor projection and a decrease in the trajectory density in four central cells.

A further area of research is in the creation of a hardware and software tool for monitoring and correcting emotional responses and cognitive activity.

References

1. 2 Shu, L., Xie, J., Yang, M., Li, Z., Li, Z., Liao, D., Xu, X., Yang, X.: A Review of Emotion Recognition Using Physiological Signals. Sensors 18(7), 2074 (2018). doi: 10.3390/s18072074.
2. 3 Panischeva S.N., Panischev O.Yu., Demin S.A., Latypov R.R.: Collective effects in human EEGs at cognitive activity. Journal of Physics: Conference Series 1038, 012025 (2018). doi: 10.1088/1742-6596/1038/1/012025.
3. 4 Montgomery R.W., Montgomery L.D.: EEG monitoring of cognitive performance. Physical Medicine and Rehabilitation Research 3(4), 1-5 (2018). doi: 10.15761/PMRR.1000178
4. 5 Magosso E., De Crescenzio F., Ricci G., Piastra S., Ursino M.: EEG Alpha Power Is Modulated by Attentional Changes during Cognitive Tasks and Virtual Reality Immersion. Computational intelligence and neuroscience, 7051079 (2019). doi: 10.1155/2019/7051079.
5. 6 Friedman N., Fekete T., Gal K., Shriki O.: EEG-Based Prediction of Cognitive Load in Intelligence Tests. Frontiers in Human Neuroscience 13(191), 1-9 (2019). doi: 10.3389/fnhum.2019.00191.
6. 1 Grissmann, S., Faller, J., Scharinger, C., Spuler, M., Gerjets, P.: Electroencephalography based analysis of working memory load and affective valence in an n-back task with emotional stimuli. Frontiers in Human Neuroscience 11(616), 1–12 (2017). doi: 10.3389/fnhum.2017.00616.
7. Perdiz, J., Pires, G., Nunes, U. J.: Emotional state detection based on EMG and EOG biosignals: A short survey. In: Proceedings of 5th Portuguese Meeting on Bioengineering (ENBENG), pp. 1-4. IEEE. Coimbra (2017). doi: 10.1109/ENBENG.2017.7889451.
8. Lee, M., Cho, Y., Lee, Y., Pae, D., Lim, M., Kang, T.: PPG and EMG Based Emotion Recognition using Convolutional Neural Network. In: Proceedings of the 16th International Conference on Informatics in Control, Automation and Robotics (ICINCO), vol. 1, pp. 595-600. Prague (2019). doi: 10.5220/0007797005950600.
9. Hsu Y.-F., Xu W., Parviainen T., Hämäläinen J.A.: Context-dependent minimization of prediction errors involves temporal-frontal activation. NeuroImage, 207, 116355 (2020). doi: 10.1016/j.neuroimage.2019.116355.
10. Ouyang G., Hildebrandt A., Schmitz F., Herrmann C.S.: Decomposing alpha and 1/f brain activities reveals their differential associations with cognitive processing speed. NeuroImage, 205, 116304 (2020). doi: 10.1016/j.neuroimage.2019.116304.
11. Duprez J., Gulbinaite R., Cohen M.X.: Midfrontal theta phase coordinates behaviorally relevant brain computations during cognitive control. NeuroImage, 207, 116340 (2020). doi: 10.1016/j.neuroimage.2019.116340.
12. Filatova, N.N., Bodrina, N.I., Sidorov, K.V., Shemaev, P.D.: Organization of Information Support for a Bioengineering System of Emotional Response Research. In: Proceedings of the XX International Conference "Data Analytics and Management in Data Intensive Domains" DAMDID/RCDL. CEUR Workshop Proceedings, pp. 90-97. CEUR. Moscow, Russia (2018) http://ceur-ws.org/Vol-2277/paper18.pdf.

13. Sidorov, K.V., Filatova, N.N., Shemaev, P.D.: An interpreter of a human emotional state based on a neural-like hierarchical structure. In: Abraham, A., et al. (eds.) Proc. of the Third International Scientific Conference "Intelligent Information Technologies for Industry" (IITI'18), Advances in Intelligent Systems and Computing, vol. 874, pp. 483–492. Springer, Switzerland (2019). doi: 10.1007/978-3-030-01818-4_48.
14. Filatova, N.N., Sidorov, K.V., Shemaev, P.D., Iliasov, L.V.: Monitoring attractor characteristics as a method of objective estimation of testee's emotional state. Journal of Engineering and Applied Sciences 12, 9164–9175 (2017). doi: 10.3923/jeasci.2017.9164.9175.

Finding the TMS-targeted group of fibers reconstructed from diffusion MRI data * (Extended Abstract)

Sofya Kulikova[0000−0002−7079−1018] and Aleksey Buzmakov[0000−0002−9317−8785]

National Research University Higher School of Economics, Perm, Russia
SPKulikova@hse.ru

Abstract. Transcranial magnetic stimulation is a promising diagnostic and therapeutic approach. Theoretical models suggest that TMS effects are local and depend on the orientation of the stimulated nervous fibers. Using diffusion MRI it is possible to estimate local orientation of the nervous fibers and to compute effects that TMS impose on them. These effects may be correlated with the experimentally observed TMS effects. However, such relationships are likely to be observed only for a small subset of the reconstructed fibers. In this work we present an approach for finding such a TMS-targeted subset of fibers within a cortico-spinal tract following stimulation of the motor cortex. Finding such TMS-targeted groups of fibers is an important for both (1) better understanding of the TMS mechanisms and (2) development of future optimization strategies for TMS-based therapeutic approaches.

Keywords: TMS · diffusion MRI · target fibers.

1 Introduction

Transcranial magnetic stimulation (TMS) is a powerful diagnostic and therapeutic approach [3]. At the whole organism level, TMS effects may be observed in various forms depending on the site of stimulation and other stimulation parameters. Theoretical models suggest that TMS effects are local and depend on the geometry of the stimulated nervous fibers [2]. Thus, the inter-subject variability of the TMS effects may arise from individual differences in the anatomy. Diffusion MRI [4] made possible 3D-reconstructions of the nervous fibers, represented by streamlines, and could be used to compute and investigate theoretically-expected TMS effects.

Combining experimental TMS data with computed TMS effects may challenge the theoretical models and give new insights on the TMS mechanisms. Since TMS effects are local, its effects should be related only to a small subset of the streamlines. Finding such a subset of streamlines is related to subgroup discovery task [1]. However, there is no subgroup discovery method searching for the best subgroup w.r.t. correlation. Thus, here we define a special brute force approach for finding a TMS-targeted group of the reconstructed streamlines.

* Supported by Russian Science Foundation grant №18-75-00034

2 Methods and Materials

2.1 Data description and pre-processing

We used TMS and MRI data of 1 healthy subject from Novikov et al. [5]. MRI data included a T1-weighted (T1w) image to create the brain conductivity model and diffusion MRI data for reconstruction of the streamlines. Data from the TMS sessions contained motor responses from the right hand and the corresponding positions of the stimulation coil. Two TMS sessions were performed on different days (Day 1 and Day 2) with the same stimulation sites.

The coil positions and the brain conductivity model were used to model the induced electric field. The effective field along the fibers was computed based on streamline geometry and the surrounding electric field. Every streamline was associate with the maximal value of the effective field, which was used to decide whether the streamline is activated or not. The search for the targeted group of fibers was done by iterating through the positions of the streamlines and through the activation thresholds maximizing the correlation between the number of activated streamlines and the amplitude of the motor response.

A brain conductivity model was computed from a T1w image using an mri2mesh algorithm [8]. Then, it was imported to SimNIBS [7] to simulate the induced electric fields for each of the stimulation sites. The results were stored as a head mesh with electric field vector assigned to each node.

Tractography and FA maps were computed according to a Diffusion Tensor Model [4] using Diffusion Toolkit[1] and the cortico-spinal tract (CST) was extracted using the regions of interest. MRI data were co-registered with TMS data using affine transformation, maximizing the mutual information between T1w image and FA maps.

2.2 Calculation of the TMS-induced effects

TMS-induced effects were computed using an in-house developed software[2]. According to [2], TMS effects depend on the effective electrical field E_l and have 3 components. The first follows from the cable equation and its effect equals to $-\lambda^2 \frac{\partial E_l}{\partial l}$ (λ is a constant, equal to 2 mm). The second is proportional to the effective field $-\lambda E_l$ and occurs at sharp bends and terminations of the fibers. The third happens at the interface between different tissue types and is at least an order of magnitude smaller that the two other effects [2]. Thus, here we considered only the first two effects. To get an electric field vector E at each point, a tetrahedra with that point is found in the head mesh and the electric field vector at the point of interest is obtained by linear interpolation from the nodes of the tetrahedra. Then, E is projected to the direction of the streamline $\boldsymbol{l}$ to get the effective field E_l. The directional derivative of the effective field is calculated as $\frac{\partial E_l}{\partial l} = \nabla E_l \cdot \boldsymbol{l}$. The total effect is obtained as a sum of the two effects.

[1] http://www.trackvis.org

[2] https://github.com/KulikovaSofya/StimVis_TMS

2.3 Finding TMS-targeted groups of fibers

Assuming that the streamlines responsible for similar behaviour are closely located, the task is to find closely placed streamlines, such that their activation correlates with the registered motor response. One activation hypothesis is that if the maximal electric field induced on the streamline is higher than a threshold, the streamline activates. However, the threshold is unknown. Thus, we need to find both a good set of streamlines and the activation threshold. The activation level for the group of streamlines is measured as the number of the activated streamlines. Thus, to get how well the selected group of streamlines is related to the motor response we correlate the number of the activated streamlines in the group with the amplitude of the motor response.

The search for the best subgroup is formalized as following: find the best rectangle in the horizontal plain and the best threshold, such that the number of activated streamlines having the coordinate within the rectangle is highly correlated with the motor response. This task can be solved by a brute force approach. However, to make it computable, some simplification are required. First, we put the coordinate system in the horizontal plane such that x-axis points from the left to the right and the y-axis is in the caudo-rostral direction. Then, we are interested only in the rectangles that are parallel to the axes. Finally, for every rectangle we allow only 10 possible positions for every edge, making it 10^4 possible rectangles in total.

To find the best threshold, we compared the induced electric field in the left and right CST (only the left CST was stimulated). Thus, the first threshold was selected to be higher the induced electric field in the right CST. Then, all higher values for the left CST were divided into 5 equally-sized groups.

Such a brute force approach introduce multiple hypothesis testing, thus, the result were validated on an independent data from experimental Day 2 (same coil positions, same subject). Accordingly, the found group of streamlines and the corresponding threshold are considered as the only hypothesis. Thus, for the new TMS positions the electric field was modeled. Then, for the previously found group of streamlines the number of the activated streamlines was computed. This number was correlated with the newly recorded motor responses.

3 Results

The reconstructed CST contained 1358 streamlines, almost equally splitted between the hemispheres. The target hemisphere demonstrated higher induced electric field and the level of 1.5mV was further used as the activation threshold.

The brute force search was applied to the TMS data of the Day 1. The best found group (12 streamlines) corresponded to the minimal activation threshold of 1.5mV. Correlation between the number of activated streamlines and the motor response was 0.51 ($p = 6 \cdot 10^{-5}$). This result was validated on the Day 2 for the same set of streamlines with the correlation of 0.37 ($p = 0.5 \cdot 10^{-}2$).

4 Discussion and Conclusions

This study presents the first attempt to test the relationships between the theoretical models of TMS mechanisms and the experimental TMS effects. In agreement with our expectations, the relations between observed and computed effects were revealed only for a small subset of the fibers from the right motor cortex. This group was stable across two independent sessions, justifying the obtained result. Since each streamline corresponds to a set of nervous fibers, it is possible to determine where the TMS-targeted fibers are located. Of importance, our results suggest that closely located fibers are responsible for similar behaviour.

However, the absolute correlation coefficients were not high. This may be related to several factors, including differences in the stimulation coils in the experiments and in the SimNIBS simulations, errors in the estimation of the fiber directions during tractography and undergoing non-stationary neuronal activity [6]. Nevertheless, the proposed approach could reveal a relevant group of fibers with a reliable relationship to the observed effects.

References

1. Atzmueller, M.: Subgroup discovery. Wiley Interdisc. Rew.: Data Mining and Knowledge Discovery **5**(1), 35–49 (2015), https://doi.org/10.1002/widm.1144
2. Geeter, N.D., Crevecoeur, G., Leemans, A., Dupré, L.: Effective electric fields along realistic DTI-based neural trajectories for modelling the stimulation mechanisms of TMS. Physics in Medicine and Biology **60**(2), 453–471 (dec 2014). https://doi.org/10.1088/0031-9155/60/2/453
3. Iglesias, A.: Transcranial Magnetic Stimulation as Treatment in Multiple Neurologic Conditions. Curr Neurol Neurosci Rep **1**(20) (2020). https://doi.org/10.1007/s11910-020-1021-0
4. Le Bihan, D., Mangin, J.F., Poupon, C., Clark, C.A., Pappata, S., Molko, N., Chabriat, H.: Diffusion tensor imaging: Concepts and applications. Journal of Magnetic Resonance Imaging **13**(4), 534–546 (2001). https://doi.org/10.1002/jmri.1076
5. Novikov, P., Nazarova, M., Nikulin, V.: Tmsmap - software for quantitative analysis of tms mapping results. Frontiers in human neuroscience **12**(239) (2018). https://doi.org/10.3389/fnhum.2018.00239
6. Peters, J.C., Reithler, J., de Graaf, T.A., Schuhmann, T., Goebel, R., Sack, A.T.: Concurrent human tms-eeg-fmri enables monitoring of oscillatory brain state-dependent gating of cortico-subcortical network activity. Commun Biol **3**(40), 1176–1185 (2020)
7. Saturnino, G.B., Madsen, K.H., Thielscher, A.: Electric field simulations for transcranial brain stimulation using FEM: an efficient implementation and error analysis. Journal of neural engineering **16**(6), 066032 (2019). https://doi.org/10.1088/1741-2552/ab41ba
8. Windhoff, M., Opitz, A., Thielscher, A.: Electric field calculations in brain stimulation based on finite elements: An optimized processing pipeline for the generation and usage of accurate individual head models. Human Brain Mapping **34**(4), 923–935 (2013)

Building models for predicting mortality after myocardial infarction in conditions of unbalanced classes, including the influence of weather conditions (Extended Abstract)

I. L. Kashirina [0000-0002-8664-9817] and M. A. Firyulina [0000-0003-3468-5514]

Voronezh State University, 1 Universitetskaya pl., Voronezh, 394006, Russia
mashafiryulina@mail.ru

1 Introduction

The key direction in modern medicine is the development of software systems that allow us to analyze a large amount of data in an adaptive way and interpret the results obtained, ensuring high accuracy of results. Predicting mortality from myocardial infarction (MI) and identifying significant factors influencing this mortality is an urgent task, since the share of cardiovascular diseases annually accounts for more deaths than any other cause [1].

The purpose of this study is to develop a machine learning model based on the gradient boosting technique for predicting mortality from MI, identifying the most significant signs, assessing the influence of meteorological factors, and improving the accuracy of predicting using data balancing methods.

In recent years, many works have been published devoted to predicting mortality from cardiovascular diseases, including those using gradient boosting methods [2,3,4]. The difference in this study lies in the addition of climatic indicators, as well as in a practical study of the effectiveness of class balancing methods in relation to the available source data. To build the model, one of the popular methods was chosen – gradient boosting, which is highly accurate

During processing real data, there are often situations when the share of examples of one class in the training dataset is too small (a minority class), and another is strongly represented (a majority class), the problem of class imbalance can lead to a serious shift towards the majority class, degradation of predicting efficiency and an increase the number of false predictions. One of the approaches to solving this problem is the use of various sampling strategies. This article discusses the various undersampling methods and compares their accuracy.

2 Materials and methods

For the analysis, we used depersonalized data on all patients who were admitted to Voronezh region's hospitals in 2015-2017 with a diagnosis of MI, and dependencies

in the sample with fatal cases of myocardial infarction (MI) for the same years. The source file contained information on 15 attributes presented in Table 1. The analyzed dataset was supplemented with six meteorological indicators. Weather data were downloaded from the rp5.ru website archive. Data preprocessing was performed using the Oracle SQL developer integrated development environment in the SQL language.

Table 1. Initial attributes.

Attribute's type	Attribute
Categorical variable	Gender, whether the myocardial infarction is repeated (MI), localization, KILLIP class, whether the patient underwent thrombolytic therapy (TLT), percutaneous coronary intervention (PCI), whether the patient has a history of diabetes mellitus (DM), atrial fibrillation (AF), acute cerebral circulation disorder (CCD), chronic obstructive pulmonary disease (COPD), chronic cardiovascular failure (CHF), arterial hypertension (AH)
Continuous variable	Age, Maximum of the temperature (Max_T), humidity, atmospheric pressure, wind speed, cloudiness

The proportion of surviving patients is much higher, which indicates the specificity of the problem being solved - the imbalance of the initial sample. For medical's tasks, the stage of balancing data is important. Rebalancing can be done in two ways: undersampling and oversampling. Oversampling is a duplication of minority class examples. Undersampling is the removal of majority class examples.

Undersampling is considered the simplest and at the same time the most correct in the tasks of medical research. Therefore, it was this method that was chosen to solve the problem. The undersampling technique can be implemented in several ways. To achieve the highest prediction accuracy, five algorithms were considered, and the accuracy of their work was compared [4].

The gradient boosting machine learning model was used to predict the mortality of patients from MI. Five-fold cross-validation was used to train the model to correct hyperparameters, the final testing was carried out on a deferred validation set, the volume of which is 20% of the initial data. Gradient boosting is a machine learning technique, the main idea of which is the iterative process of sequentially building partial models. Each new model is trained using information about the errors made at the previous stage, and the resulting function is a linear combination of the entire ensemble of models, considering the minimization of some penalty function [5]. This algorithm is distinguished by its high accuracy, in most cases, surpassing the accuracy of other methods.

In machine learning tasks, various metrics are used to assess the quality of models. When calculating these metrics, a classification error matrix is used. The main metric of classification problems is the proportion of correct answers. Medical diagnostics has its own metrics that determine the accuracy of the method [6]. Sensitivity (true positive proportion) reflects the proportion of positive results that are correctly identified as such and the ratio of the number of deaths classified as deaths to the total number of deaths is calculated. Specificity (true negative proportion) reflects the proportion of negative

results that are correctly identified as such and the ratio of the number of survivors classified as survivors to the total number of survivors is calculated.

3 Results and discussion

The gradient boosting model was built using the XGBClassifier library tools. The data was divided into test and training samples in a 20:80 ratio. The model accuracy was 0.85, the sensitivity and specificity indices were 0.35 and 0.97, respectively.

The significance of each factor is calculated as the average normalized result of the decrease in the branching criterion caused by this factor. The branching criterion calculates the measure of uncertainty at the nodes of the trees. The Gini index was used as such a criterion. The most significant sign is the Killip severity score. Age is expected to be a significant factor; survival is worse in patients belonging to the older age group. Also, predictors influencing the results include indicators: whether percutaneous coronary interventions (PCI) were performed, whether the patient has chronic heart failure (CHF), has a history of stroke (CCD) and chronic obstructive pulmonary disease (COPD). It can be noted that in comparison with clinical indicators, meteorological factors are less significant, however, there is an influence of cloudiness indicators (Cloudiness) and maximum daily temperature (Max_T).

The undersampling method is used to balance the classes. The considered strategies and their quality metrics are presented in the Table 2. Based on the results of the model quality metrics, we can conclude that the best indicators of accuracy when using strategies of random deletion, cluster centroids and Tomek links. In this case, the method of cluster centroids provides the highest sensitivity.

Table 2. Quality metrics for various sampling strategies.

Method	Accuracy (train)	Sensitivity	Specificity	Accuracy (test)	AUC_ROC
Random undersampling	0.79	0.50	0.91	0.83	0.81
Cluster centriods undersampling	0.85	0.68	0.75	0.74	0.79
TomekLinks undersampling	0.85	0.40	0.98	0.86	0.82
NearMiss undersampling	0.80	0.48	0.88	0.81	0.76
Scale pos weight	0.83	0.16	0.99	0.83	0.80

4 Conclusion

This study was conducted to build a gradient boosting model for predicting mortality after myocardial infarction and to determine the most significant factors for mortality in MI. The accuracy of the model was improved by balancing the original sample using undresampling. The accuracy of five sampling strategies has been demonstrated. The most effective methods are the method of random removal of samples, the method of cluster centroids and the method of Tomek links. The best accuracy (Accu-

racy) was obtained using the Tomek links method (on test data, it is equal to 0.85). The best sensitivity is provided by the cluster centroid method.

With the help of the constructed model, significant factors were found for predicting mortality after the onset of myocardial infarction. The most significant are: Killip class, age, percutaneous coronary interventions, whether the patient has chronic heart failure, whether there is a history of stroke and chronic obstructive pulmonary disease, as well as weather factors – cloudiness and maxi-mum daily temperature.

References

1. Heron, M.: Deaths: Leading causes for 2017. National Vital Statistics Reports (6), 1–96 (2018).
2. Weng SF, Reps J, Kai J.: Can machine-learning improve cardiovascular risk prediction using routine clinical data. PLoS One. 12(4) (2017).
   Ahmad T, Lund LH, Rao P.: Machine Learning Methods Improve Prognostication, Identify Clinically Distinct Phenotypes, and Detect Heterogeneity in Response to Therapy in a Large Cohort of Heart Failure Patients. Journal of the American Heart Association. 7(8) (2018).
3. Ambale-Venkatesh B, Yang X, Wu CO.: Cardiovascular Event Prediction by Machine Learning: The Multi-Ethnic Study of Atherosclerosis. Circ Res. 121(9) (2017).
4. Garcia S., Herrera F.: Evolutionary Undersampling for Classification with Imbalanced Datasets: Proposals and Taxonomy. Evolutionary Computation 17(3), 275–306 (2009).
5. Brownlee, J.: XGBoost With Python: Gradient Boosted Trees with XGBoost and scikit-learn. 2nd edn. Machine Learning Mastery Pty. Ltd, (2018).
6. Glazkova, T.: Assessment of the quality of diagnostic methods and prognosis in medicine. Bulletin of science center of medical Sciences of Russia (2), 3–11 (1994).

Renal Impairment Risk Factors in Patients with Type 2 Diabetes (Extended Abstract)

D. A. Shipilova [0000-0003-4862-6208], O. A. Nagibovich [0000-0002-1520-0860]

Military Medical Academy named after S.M. Kirov,
Saint-Petersburg, Akademika Lebedeva street, 6G, Russia
dashuta_shipilova@mail.ru

1 Introduction

Chronic kidney disease (CKD) is the leading microvascular complication in patients with type 2 diabetes mellitus (DM) [1, 2]. Currently, the diagnosis of diabetic kidney damage is mainly based on clinical and laboratory parameters. According to the recommendations of national and foreign expert groups to clarify the stage of kidney damage in diabetes, it is necessary to determine the glomerular filtration rate (GFR) and the level of albumin excretion in the urine [3, 4]. These indicators characterize both structural and functional changes in the kidneys, but they do not predict the rate of progression of diabetic kidney damage. It is known that the risk factors for the development and progression of deterioration of renal function are unsatisfactory compensation of diabetes mellitus, long duration of the disease, impaired intrarenal hemodynamics, arterial hypertension, hyperlipidemia [5]. However, it is still difficult to predict the renal prognosis, since there are additional causes of the loss of renal filtration capacity. So, according to some authors, Doppler ultrasound, namely, the determination of the index of intrarenal vascular resistance, can detect hemodynamic disturbances in the early stages of renal disease and prevent the onset or progression of renal failure in patients with diabetes [6-8]. In addition, R. Ikee et al found a pronounced relationship between histopathological parameters and the level of the resistance index (RI) in patients with type 2 diabetes [9]. H. Xu et al., in 2017, conducted a study in which, using renal Doppler sonography, revealed that microvascular disorders are an early marker of nephrosclerosis, even before the detection of morphological changes in mice with induced type 2 diabetes [10]. Thus, the search for additional risk factors will expand the understanding of the pathogenesis of diabetic kidney damage, and will allow developing effective approaches to prevent the progression of CKD.

Purpose of the study is to identify risk factors for deteriorating renal function and determine the possibility of using RI as a criterion for predicting renal outcome in patients with type 2 diabetes.

2 Materials and methods

From October 2015 to December 2019, 82 patients with type 2 diabetes were under observation. During the observation period, 26,8% of patients showed a decrease in GFR below 60 ml/min/1,73 m2 and an increase in the albumin-creatinine ratio (A/Cr) above 3 mg/mmol. In this regard, the patients were divided into two groups: 1st - 60 patients (39 men and 21 women), who had a GFR higher or equal to 60 ml/min/1,73 m2 and A/Cr from 0,2 to 1,8 mg/mmol. 2nd - 22 patients (18 men and 4 women), who had GFR below 60 ml/min/1,73 m2 and A/Cr from 5 to 16,4 mg/mmol. The main clinical and laboratory parameters were studied: anthropometric, hemodynamic, level of carbohydrate metabolism compensation, serum creatinine, urine creatinine, urine albumin, lipid metabolism indicators. All patients underwent a Doppler study of one of the segmental arteries of the right kidney. The index RI of intrarenal vascular resistance was determined. The diagnosis of CKD was established based on the determination of GFR using the CKD-EPI formula and the calculation of A/Cr in accordance with the recommendations of the International Society of Nephrology (KDIGO). Statistical processing of the experimental data was carried out using the STATISTICA 10 software and included methods of variation statistics, correlation analysis, and nonparametric tests (Mann-Whitney, Ro-Spearman) [11]. The analysis of factors affecting renal outcome was carried out in several stages. At the first stage, a statistically significant difference in the studied parameters was revealed. Threshold values were generated for each selected factor. At the second stage, a one-way analysis of renal outcome according to Kaplan-Meier using a long-rank test was performed [12]. At the final stage, multivariate analysis was carried out using the Cox regression model. The magnitude of the relative risk was determined with the indication of 95% confidence interval (CI). Sample data are presented in the form Me [Xmin; Xmax], where Me is the median of the sample data, [Xmin; Xmax] is the range of the sample. In paired comparisons, the level of significance $\alpha = 0,05$ is accepted.

3 Results and discussion

The study revealed that patients with type 2 diabetes with signs of CKD (group 2) were older compared to group 1 (65 [60;70] vs 56[50;63] years, p2,1=5,5*10-8, respectively). In addition, they differed in a longer duration of the disease (15[10;23] vs 6[5;8] years, p2,1=0,0001, respectively). The presence of obesity (31[29;33] vs 29[26;32] kg/m2, p2,1=0,009, respectively); a higher level of glycated hemoglobin (HbA1c) (9,0[7,8;10,0] vs 7,9[6,9;8,9] %, p2,1=0,002, respectively) and a higher RI value (0,71[0,67;0,73] vs 0,66[0,61;0,69], p2,1=0,000001, respectively).

In the general group of the surveyed, a relationship was found between serum creatinine and: age (ρ=0,38, p=0,0001), the duration of diabetes (ρ=0,35, p=0,0004), BMI (ρ=0,28, p=0,005), RI (ρ=0,41, p=0,0001). And also between GFR and: age (ρ=-0,55, p=0,002), BMI (ρ=-0,30, p=0,0001), HbA1c (ρ=-0,32, p=0,001), RI (ρ=-0,38,

p=0,001), relationship between A/Cr and: RI (ρ=0,30, p=0,001), HbA1c (ρ=0,31, p=0,001).

For univariate analysis of renal outcome using the Kaplan-Meier method, patients were divided into two groups depending on the duration of diabetes mellitus: duration of diabetes <10 years and duration of diabetes ≥10 years. This threshold value was chosen because it is known that the development of irreversible morphological changes begins 10 years after the onset of DM [13]. Deterioration of renal function during the follow-up period was found equal to 40% in the group with DM duration ≥10 years (p = 0,0001).

For BMI, the upper limit of 30 kg/m2 is taken as the threshold value in accordance with the standards of the World Health Organization (WHO). Two groups of patients were identified. In the group with BMI ≥30 kg/m2, the deterioration of renal function was 45% (p = 0,0001).

For HbA1c, the threshold was 8% based on algorithms for specialized medical care for patients with diabetes [3]. Two groups of patients were identified. Deterioration of renal function during the observation period was detected in the group with HbA1c ≥8% and amounted to 45% (p = 0,001).

For age, the upper limit of the norm of 60 years (elderly age) was taken as a threshold value and 2 groups of patients were identified. In the group of patients aged ≥60 years, the deterioration of renal function was 50% (p = 0.0001).

It is known that an increase in the resistance index values above 0,70 indicates a decrease in renal function in patients with type 2 DM [6]. Patients by RI were divided into two groups: RI <0,70 and RI ≥ 0,70. Deterioration of renal function over the observation period was found in the group with RI ≥ 0,70 and amounted to 60% (p = 0,001).

At the final stage, a Cox model was formed, which was based on all factors affecting the renal outcome identified at the previous stages: age, duration of the disease, BMI, HbA1c, RI. The analysis showed that a significant risk factor for deterioration of renal function is RI≥0,70 (Er=1,9; CI=1,6–2,3; p=0,001). In particular, patients with a high RI had a 1,9 times greater risk of death than patients with normal values.

In conclusion, would like to note that the complexity of the problem we are solving required a more sophisticated method of statistical analysis. Along with analysis of variance, logistic regression, survival analysis takes a significant place. This is a method by which, over a certain period of time, the patterns of the appearance of a certain outcome in representatives of the observed sample are studied. There are several mathematical and statistical methods that can be used to analyze survival, in cases where there is incomplete information about the sample: using life tables, the Kaplan-Meier method, Cox regression and Cox regression with time-dependent predictors [14]. In this article, we examined the possibility of using the Kaplan-Meier method and Cox regression on a real example of analyzing risk factors affecting renal outcome in patients with type 2 DM.

4 Conclusions

Thus, using the Kaplan-Meier method and Cox regression, we confirmed that the risk factors for deteriorating renal function in patients with type 2 diabetes are:

- age,
- duration of diabetes,
- obesity,
- level of compensation of carbohydrate metabolism,
- index of resistivity.

According to the results of multivariate analysis in the Cox model, the most significant factor for predicting renal outcome is a resistance index equal to 0,70 and higher. The definition of this indicator can be used for non-invasive diagnosis and assessment of kidney damage in patients with type 2 diabetes.

References

1. Shamkhalova, M.S. Trends in the epidemiology of chronic kidney disease in Russian Federation according to the Federal Diabetes Register (2013–2016). Diabetes Mellitus 21(3), 160-169 (2018).
2. Klimontov, V.V. Cystatin C and collagen type IV in diagnostics of chronic kidney disease in type 2 diabetic patients. Diabetes Mellitus 18(1), 87-93 (2015).
3. Dedov, I.I. Standards of specialized diabetes care. Diabetes Mellitus 22(S1), 68-79 (2019).
4. Shilov, E.M., Smirnov, A.V., Kozlovskaya, N.L. Nephrology. Clinical guidelines. GEOTAR-MEDIA, Moscow (2020).
5. Shestakova, M.V., Martynov, S.A. Diabetic Nephropathy: Do We Consider All Risk Factors? Diabetes Mellitus 4, 29-33 (2006).
6. Mancini, M.: Renal duplex sonographic evaluation of type 2 diabetic patients. Journal of Ultrasound in Medicine 32(6), 1033–1040 (2013).
7. Kopel, J., Pena-Hernandez, C, Nugent, K.: Evolving spectrum of diabetic nephropathy. World journal of diabetes 10(5), 269 (2019).
8. Kim, J.H.: Resistive index as a predictor of renal progression in patients with moderate renal dysfunction regardless of angiotensin converting enzyme inhibitor or angiotensin receptor antagonist medication. Kidney research and clinical practice 36(1), 58–67 (2017).
9. Ikee, R. Correlation between the resistive index by Doppler ultrasound and kidney function and histology. American journal of kidney diseases 46(4), 603–609 (2005).
10. Xu, H. Renal resistive index as a novel indicator for renal complications in high-fat diet-fed mice. Kidney and Blood Pressure Research 42(6), 1128–1140 (2017).
11. Trukhacheva, N.V. Mathematical statistics in biomedical research using the Statistica package. GEOTAR-MEDIA, Moscow (2013).
12. Glantz, S. Medical and biological statistics. Practice, Moscow (1999).
13. Nadeeva, R.A., Sigitova, O.N. Clinical guidelines for the treatment of diabetic nephropathy. Archive of internal diseases 5, 3-8 (2015).
14. Sharashova, E.E. Survival analysis in health sciences using SPSS software. Science & Healthcare 5-28 (2017).

Методы и средства анализа сигналов головного мозга человека на данных функциональной магнитно-резонансной томографии (Расширенные тезисы)

Д. Ю. Ковалев[1], Д. И. Сергеев[2], Е. М. Тириков[2], Н. В. Пономарева[3]

[1] Институт проблем информатики ФИЦ ИУ РАН, Москва
[2] Московский государственный университет им. М. В. Ломоносова, Москва
[3] Федеральное государственное бюджетное научное учреждение "Научный центр неврологии", Москва

dm.kovalev@gmail.com em.tirikov@gmail.com
serdimigor@gmail.com ponomare@yandex.ru

1 Введение

Нейроинформатика находится на стыке нейрофизиологии и информатики и является междисциплинарной наукой, изучающей методы и инструменты анализа деятельности человеческого мозга. Проблема анализа заключается не только в объеме накопленных данных, но и в различных типах и структуре данных. В частности, данные фМРТ представляют собой 4-х мерные изображения (одна временная координата и три пространственных). Различают два типа фМРТ: 1) изображения фМРТ в состоянии покоя; 2) изображения фМРТ действия. Среди основных типов взаимодействия между регионами мозга выделяют эффективное и функциональное. Анализ функциональной и эффективной связности позволяет выделять как нейросети покоя, так и оценить участие конкретных структур мозга в обеспечении сложных функций, таких как язык, память и пр. Целью данной работы является обзор существующих методов и средств анализа сигналов головного мозга человека для данных фМРТ, выбор задач для распределенной реализации, а также сравнение решения приведенных задач с родственными работами. В разделе 2 формулируются постановки задач анализа сигналов фМРТ. В разделе 3 приведены существующие методы анализа сигналов фМРТ. В разделе 4 приводятся результаты. Работа выполнена при поддержке Российского Фонда Фундаментальных Исследований, идентификатор проекта 18-29-22096 мк.

2 Задачи по анализу сигналов фМРТ в нейрофизиологии

2.1 Поиск значимых различий нелинейной функциональной связности головного мозга для мужчин и женщин в состоянии покоя

Существует два типа функциональной связности головного мозга человека: линейная и нелинейная. В большинстве случаев исследуется лишь линейная функциональная связность [1]. В работе [2] показано, что функциональная связность между некоторыми областями мозга является нелинейной, а одним из способов изучения нелинейной функциональной связности является построение и исследование аналитических функций с помощью генетического программирования.

Задача поиска различий в работе головного мозга между мужчиной и женщиной уже давно интересует нейрофизиологов. В статье [3] показано, что у мужчин и женщин в состоянии покоя есть различия показателей фМРТ в первичной зрительной коре, задней срединной префронтальной сети и других отделов головного мозга. В работе [4] демонстрируется, что при рассеянном склерозе у мужчин проявляется более слабая активность в хвостатом ядре по сравнению с женщинами. Недостатком рассмотренных выше работ является применение лишь линейных методов для сравнения. При использовании нелинейной функциональной связности становится возможным более подробно исследовать взаимосвязь между регионами головного мозга и лучше понять его работу в целом.

2.2 Определение активности мозга человека с использованием функциональной магнитно-резонансной томографии действия

Классификация активности мозга человека с использованием фМРТ действия является важной частью анализа эффективной связи. Например, в проекте Human Connectome Project (HCP) [5] участников попросили выполнить семь задач, относящихся к следующим категориям: эмоции, азартные игры, язык, социальное взаимодействие, двигательная активность, реляционная, и рабочая память:

Каждый участник эксперимента имеет соответствующие данные возраста, веса, пола и т. д. Данные фМРТ по каждому участнику хранятся в отдельном файле проекта, в котором хранится метаинформация об эксперименте, количество и название записанных условий эксперимента, а также указание пути к временным файлам. Временные файлы содержат время стимула (начало) и его продолжительность.

3 Методы и средства анализа сигналов головного мозга человека

3.1 Выделение регионов головного мозга из фМРТ изображений

Извлечение регионов головного мозга является общим этапом для обеих приведенных задач. Не существует универсального способа извлечения регионов головного мозга [6]. Методы извлечения регионов для одного человека подразделяют анализ поверхности головного мозга каждого субъекта независимо друг от друга. Часть методов в данном подходе основана на широко используемых алгоритмах кластеризации. При работе с атласами головного мозга человека используется заранее вычисленная функция, ставящая каждому вокселю изображения в соответствие регион головного мозга.

3.2 Методы построения нелинейных аналитических функций для анализа функциональной связности

Обобщенная линейная модель с нелинейными членами является одним из популярных методов изучения нейрофизиологических изображений [1], т.к. она является быстро вычислимой, а полученный результат просто интерпретируется. Недостатком метода является необходимость определения нелинейные зависимости заранее. Для автоматического восстановления нелинейных функциональных связей также используется метод генетического программирования [2]. Преимуществом алгоритма является то, что заранее не делается никаких выводов о функциональной связности. Существенным недостатком является его экспоненциальная вычислительная сложность.

В качестве основного алгоритма для построения нелинейной функциональной связности выбран алгоритм генетического программирования, так как он является хорошо интерпретируемым и не требует построения дополнительных признаков. Для сравнения полученных функций для мужчин и женщин авторами была проведена проверка гипотезы с использованием рангового критерия для связных выборок и поправкой Холма для множественного тестирования.

3.3 Методы и средства построения классификаторов активности головного мозга

Работы по анализу изображений фМРТ с использованием нейронных сетей начали появляться сравнительно недавно. В работе [7] исследователи использовали глубокую нейронную сеть с пятью свёрточными и двумя полносвязными слоями. Авторам статьи удалась достичь высокой точности классификации 93.7%. В [8] предлагает использовать сверточные нейронные сети (AlexNet, Inception, ResNet) с точность классификации около 95%. В классе группирующих моделей наиболее распространёнными являются модели градиентного бустинга и их модификации [9].

4 Результаты

4.1 Результаты для задачи поиска различий нелинейной функциональной связности

На первом этапе были обработаны все данные фМРТ изображений в состоянии покоя, т. е. 4532 изображения (для каждого человека в проекте HCP имеется 4 изображения). Всего были проанализированы данные 563 женщин и 550 мужчин возраста от 20 до 35 лет. Регионы были извлечены при использовании атласа Harvard-oxford. На втором этапе производился расчет функций для 48 регионов с использованием метода генетического программирования. В качестве метрики используется коэффициент детерминации с пороговым значением. Для мужчин и женщин вычислены функции для 23 регионов. Последним этапом является проверка гипотезы о присутствии значимого различия нелинейных функций. В результате демонстрируется, что из 23 регионов существенные различия имеют 11 регионов.

4.2 Результаты для задачи классификации активности мозга человека

В наборе данных проекта HCP для данных фМРТ действия продолжительность эксперимента по каждому классу отличается, что свидетельствует об особенностях проведения эксперимента. Поэтому все данные фМРТ обрезаются по минимальной длине временного ряда. Другие особенности связаны с различными факторами: видом задачи, к примеру, при просмотре видео ролика происходили задержки по воспроизведению видео, продолжительностью каждой сессии внутри эксперимента и т.д. Для их устранения классифицируются не весь временной ряд, а случайно выбранный интервал временного ряда (100 точек — это составляет 30% процентов от исходной длины ряда).

Все группирующие модели демонстрирую схожее поведение. По результатам исследований качество классификаторов на атласе AAL в среднем выше на 1.5-2%, чем на атласе Harvard-oxford. Атласы с небольшим количеством регионов (40-50), демонстрирую высокие значение метрик качества 92% при небольшой размерности признакового пространства (менее 100 признаков). Использование атласа с большим количеством регионов (100-150) позволяет достичь качества классификации на уровне 96% за счет значительного увеличения признакового пространство (150-250).

Ссылки

1. Pierre-Jean L., Jean-Baptiste P., Guillaume F., Silke Dodelline G.: Functional connectivity: studying nonlinear, delayed interactions between BOLD signals, 20 (2), 962-974 (2003).
2. Allgaier, N., Banaschewski, T., Barker, G., Bokde, A. L., Bongard, J. C., Bromberg, U.,et al (2015). Nonlinear functional mapping of the human brain. arXiv preprint arXiv:1510.03765.

3. C. Xu, C. Li, H. Wu, etc.: Gender differences in cerebral regional homogeneity of adult healthy volunteers: a resting-state FMRI study. Biomed research international, (2015).
4. Zang Y, Jiang T, Lu Y, et al. Regional homogeneity approach to fMRI data analysis. Neuroimage, 22, 394–400, 2004.
5. David C. Van Essen, Stephen M. Smith, Deanna M. Barch, Timothy E.J. Behrens, Essa Yacoub, Kamil Ugurbil, for the WU-Minn HCP Consortium. The WU-Minn Human Connectome Project: An overview. NeuroImage 80:62-79. (2013). http://www.humanconnectomeproject.org/
6. Arslan, S., Ktena S. I., Makropoulos A., Robinson E. C., Rueckert D., Parisot S.: Human brain mapping: A systematic comparison of parcel-lation methods for the human cerebral cortex. Neuroimage 170, 5-30, (2018).
7. Xiaoxiao, W., Xiao, L., Zhoufan J., Benedictor, A., Yawen Z., Yanming, W., Huijuan, W., Yu, L., Yuying, Z., Feng, W., Jia-Hong, G., Benching, Q.: Decoding and mapping task states of the human brain via deep learning. arxiv.org (2018).
8. Yufei, G., Yameng, Z., Hailing, W., Xiaojuan, G., Jiacai, Z.: Decoding Behavior Tasks From Brain Activity Using Deep Transfer Learning.IEEE Access (2019)
9. XGBoost - https://xgboost.readthedocs.io/en/latest/

Application Association Rule Mining in Medical-biological Investigations: a Survey (Extended Abstract)

Xenia Naidenova, Vyacheslav Ganapolsky, Alexandre Yakovlev,

and Tatiana Martirova

Military Medical Academy, Saint Petersburg, Russia
E-mail: ksennaidd@gmail.com

1 Introduction

Intelligent data processing is now an integral part of biomedical research. Revealing different dependencies in the data (implicative, functional, correlational, etc.) helps in diagnosis, treatment's planning, predicting the course of diseases and in identifying new factors that expand the understanding of specialists about specific diseases and their combinations.

The purpose of many biomedical studies is to highlight associative rules in a given data set. The association rule is the rule in the form $X \Rightarrow Y$, where X and Y are non-intersecting sets of distinct literals called items. In general case, we can consider a set of items as a set of all attributes' values that can appear in descriptions of some objects or situations, consequently, a transaction in a data base as a collection of attribute values composing a description of some object or situation is an itemset.

Let $I = \{i_1, i_2, \ldots, i_N\}$ be a set of items. A transaction database (TDB) is a set of transactions, where transaction <tid, X> contains a set of items (i.e., $X \subseteq I$) and is associated with a unique transaction identifier tid. A non-empty itemset $Y \subseteq I$ is called q-itemset if it contains q items. A transaction <tid, X> is said to contain itemset Y if $Y \subseteq X$. The number of transactions in TDB containing itemset X is called the support of itemset X, denoted as sup(X): sup (X) = |{tid| (tid, Y) $\in$ TDB, $X \subseteq Y$}|, where |s| denotes the cardinality of s. Giving a minimum support threshold, min-sup, an item Y is frequent if $\sup(Y) \geq$ min-sup.

In frequent itemsets, association rules are extracted in the form of implications, for which the value of support is an important characteristic: $\text{Sup(rule)} = \sup(X \Rightarrow Y) = \sup(X \cup Y)$. The rule has a measure of reliability called confidence and defined as follows: $\text{conf}(X \Rightarrow Y) = \sup(X \cup Y)/\sup(X)$. Confidence is defined as the part of all transactions in TDB containing X and Y, among those transactions that contain X. The third main metric of association rules is Lift defined as follows: $\text{lift}(X \Rightarrow Y) = \text{conf}(X \Rightarrow Y)/\sup(Y)$. The traditional purpose of data analysis is to find all associative rules that have support and confidence above the specified minimum values.

In this paper, we consider the applications of the classical Apriori algorithm for mining association rules in biomedical data. This choice of algorithm is justified by

the following arguments: the Apriori algorithm is the most understandable and easy mastered by specialists in different applied fields, it is universal algorithm, in the sense that implicative rules containing in frequent itemsets can be extracted with the help of this algorithm [13], it is constantly improving based on many progressive techniques in association rule mining.

The idea of the classical Apriori algorithm [1] is based on the following consideration: q-itemset can be frequent if and only if all its proper sub-itemsets are frequent.

At the first step of the algorithm, all the items are considered (values of attributes, elements of transactions) and among these are separated those satisfying the condition of minimal support. Then these separated items are used to form itemsets of two items (candidates for frequency). For them, the support is calculated and those that do not meet the minimum support condition are removed. The remaining itemsets are used to form ones of three items. The process is going on iteratively, as long as it is possible to generate a new set of candidates for frequency. This Apriori algorithm uses an effective inductive method of constructing sets of the cardinality (q+1) ((q+1)-sets) from their subsets of the cardinality q (q-sets). The method of forming (q +1)-sets from q-sets and calculating their supports are the main sub-processes of the Apriori algorithm determining its computational complexity.

2 The Apriori algorithm in medical studies

The Apriori algorithm is widely used in biomedical research. These studies cover: cardiovascular disease; lung cancer; oral cancer; infectious diseases (Ebola virus); breast cancer; type 2 diabetes; Alzheimer's disease; liver cancer.

The problems solved in these studies are also varied: searching for unknown trends in heart disease; determining the nature of heart disease based on a prediction method; diagnosis (detection) of disease; predicting a patient's response to drug; early diagnosis and prevention of disease; prediction of illness's progress (course of disease) (trends); predicting the outcome of disease; identification of disease risk factors; identification of relationships between different medical operations, appointments, analyses and diagnoses; extracting diagnostic patterns (sets of features, symptoms) in electronic medical database; extracting association rules in medical data (in Electronic Medical Record Systems) and many others. The papers [2, 8, 9, 10, 11, 14] give detailed reviews of associative rule extracting from biomedical data based on Apriori algorithm and its modifications.

It should be noted that there are still very few works in the domestic literature on the use of the Apriori algoritm in biomedical research. In addition to the work of Biryukov A. and Dumansky S. [6], we can cite the article [5], which proposes a new effective algorithm AprioriScale to build association rules. The algorithm is applied to the problem of detecting children's diseases: obesity and metabolic syndrome.

3 Analysis of biological and genetic data based on association rule extracting

A topic dealing with the analysis of patient biological data is now becoming particularly relevant. The biological data analysis is connected with the identification of previously unknown hidden patterns (frequent itemsets), associative structures in the large number of biological sequences. These sequences include gene sequences, amino acid sequences, protein composition, and other data that display the structure, localization, interaction or functioning of proteins and genes in cells. Amino acids are the building material of proteins. The shape and other properties of proteins are associated with the exact sequence of amino acids contained in them. The chemical properties of amino acids determine the biological activity of proteins.

Many diseases have biological nature, such diseases as obesity, high blood cholesterol, diabetes, insomnia, arthritis, and many others. Analysis of gene information, including Apriori algorithms, helps to study the nature of disease, optimize its treatment, predict the course of disease.

An overview of some methods of extracting knowledge from biological (DNA) sequences is given in [3, 7]. The comparison of the Apriori algorithm with other algorithms in the mutation analysis is produced in [12].

4 Perfection of the Apriori algorithm

The popularity of the Apriori algorithm for medical diagnostic tasks is due to its simplicity, however, its application for large data sets requires the development of more efficient modifications in terms of reducing its computational complexity. And such work to improve this algorithm is being carried out all over the world [4]. In particular, we may be able to choose the following directions in improving the Apriori algorithm:

- Developing new algorithms;
- Data management;
- Constraint-based association rules mining;
- Incremental mode of association rules mining.

Conclusion

The paper provides an overview on mining associative rules from data in biomedical research. This review is based on a study of the work from 2013 to 2020. and shows the widespread use of the Apriori algorithm and its modifications in medicine. The review includes also the methods to improve the Apriori algorithm to mining more effective associative rules adapted for various research tasks.

References

1. Agrawal, R., Imieliński, T., and Swami, A.: Mining Association Rules Between Sets of Items in Large Databases. ACM SIGMOD Record, 22(2), 207-216 (1993). DOI: 10.1145/170036.170072
2. Altaf, W., Shahbaz, M., and Guergachi,A.: Applications of association rule mining in health informatics: A survey. Artificial Intelligence Review, 47(3), 313-340 (2017).
3. Anandhavalli, M., Ghose, M.K., and Gauthaman, K.: Association Rule Mining in Genomics. International Journal of Computer Theory and Engineering, 2(2), 1793-1802 (2010).
4. Bhende, D., Kasarker, U., and Gedam, M.: Study of various improved Apriori algorithm. IOSR Journal of Computer Science Engineering, e-ISSN:2278-0661, p. 55-58 (2016).
5. Billig, V. A., Ivanova, O.V.: Building association rules in a task of medical diagnosis. Software Products and Systems, 2(114), 146-157 (2016). (in Russian)
6. Birukov, A. P., Dumansky, S. M.: Revealing the comorbidity of professional diseases caused by pathogenic factors with the help of associative algorithms on the examples of a cohort of victims of ionizing radiation. Medicine of extreme situations, № 2, 13-24 (2016). (in Russian)
7. Das, N. N. et al.: Brief Survey on DNA Sequence Mining. International Journal of Computer Science and Mobile Computing, 2(11), 129-134 (2013).
8. Doddi, D., Marath, A., Ravi, S.S., and Torney, D.C.: Discovery of Association Rules in Medical Data. Medical Informatics and Internet in Medicine 26(1), 25-33 (2001). doi:10.1080/14639230010028786
9. Kang'ethe, S. M. and Wagacha, P. W.: Extracting Diagnosis Patterns in electronic medical records using association rule mining. International Journal of Computer Applications 108(15), 19-27 (2014).
10. Kavakiotis, I., Tzanis, G., and Vlahavas, I.: Mining frequent patterns and association rules from biological data. In: Mourad Elloumi and Albert Y. Zomaya (eds.), Biological Knowledge Discovery Handbook: Preprocessing, Mining, and Postprocessing of Biological Data, First Edition, Chapter 34, pp. 737-762. John Wiley & Sons, Inc. Published (2014).
11. Kumar, R. P. R., Jayakumar, R., and Sankaridevi, A.: Apriori-based Frequent Symptomset Association Mining in Medical Databases. International Journal of Recent Technology and Engineering (IJRTE), 7(5C), 65-68 (2019).
12. Mayilvaganan, M. and Hemalathe, R.: Performance Comparison of FSA Red and Apriori Algorithm's in mutation Analysis. International. Journal of Computer Trends and Technology (IJCTT), 17(4), 205-209 (2014).
13. Naidenova, X., Parkhomenko, V., and Shvetsov, K.: Application of a logical-combinatorial network to symbolic machine learning tasks. In: Naidenova, X., Shvetsov, K., Yakovlev, A. (eds.), Machine Learning in analysis of biomedical and socio-economic data, pp. 371-409. Saint-Petersburg, Russia: Polytech Press (2020).
14. Pazhanikumar, K. and Arumugaperumal, S.: Association Rule Mining and Medical Application: A Detailed Survey. International Journal of Computer Applications 80(17), 10-19 (2013) DOI: 10.5120/13967-1698

The Use of Machine Learning Methods to the Automated Atherosclerosis Diagnostic and Treatment System Development (Extended Abstract)

Maria Demchenko[1] and Irina Kashirina[1]

[1] Voronezh State University, Voronezh 394018, Russia
masha-vrn@yandex.ru

1 Introduction

A key aspect of successful treatment is the timely diagnosis of a disease. In particular, the symptoms of the atherosclerosis disease can be found among 17% of people between the ages of 13-19 years, and by the age of 40 years, at least one atherosclerotic lesion is present in more than 70% of patients [1]. The high prevalence of this disease, as well as the accompanying high risk of vascular damage and ischemic lesions of organs, requires the thorough research and the development of the most effective diagnostic methods and treatment policies.

The highly productive machine learning algorithms are used to solve medical diagnostic problems. One of the most popular approaches is deep learning, which is actively used, in particular, in the task of diagnostics of the life-threatening diseases, which is illustrated in [2,3]. At the same time, a significant part of the studies nowadays is devoted to the application of the effective machine learning methods to the research of cardiovascular diseases. For example, in [4,5], an approach to the diagnosis of peripheral arterial diseases using supervised machine learning methods was demonstrated.

The task of developing a system for atherosclerosis diagnostics and treatment prescriptions was set by the specialists of the Voronezh Regional Cardiological Dispensary (VOCD). This research includes the following main stages.

1. The development of the non-invasive atherosclerosis diagnostic models. The solution to this problem involves 2 main steps.
   a. Development of diagnostic models using the dataset provided by the VOCD. As a training sample, this study used data collected within a research conducted among 522 adult patients from the Bogucharsky district of the Voronezh region using multichannel volume sphygmography (MVS) – an efficient non-invasive method for atherosclerosis diagnostics [6] aimed at identifying the significant asymmetry of systolic blood pressure (SBP) on the arms or legs (ArmsIndex and LegsIndex, respectively), as well as the ankle-brachial index (ABI). In particular, studies [7-9] have been devoted to the identifying of the SBP asymmetry parameters, which also can be considered as diagnostic fea-

tures of atherosclerosis. In this study, neural network models (binary classifier of MLP architecture, self-organizing Kohonen maps) [10], as well as ensemble models (RandomForest, ExtremeRandomTrees, XGBoostClassifier) [11] were used to find the most informative values of the SBP asymmetry coefficients and use them as the basis for the calculation of the atherosclerosis markers values.

   b. Development of the diagnostic models using a sample of the international database MIMIC-III. Also at this stage of the study, the Azure Machine Learning platform was used as a tool for building diagnostic models, which allowed to develop scalable solutions with the possibility of deployment and continuous training. [12]

2. The development of the optimal patients' treatment strategies using efficient reinforcement learning models [13,14].
3. The development of an automated system for collecting medical information, including patients' diagnoses and doctors' appointments, as well as performing administrative functions (registering patients, scheduling consultations with a doctor, etc.).

2 Atherosclerosis diagnostic models and methods

2.1 The task of the atherosclerosis markers and predictors analysis. The dataset of the Voronezh Regional Cardiology Dispensary

Within this study we were able to take into account the specific properties of the atherosclerosis diagnostics problem for a group of patients using the dataset of patients from the Bogucharsky district of the Voronezh region, provided by VOCD [6,15].

As a result of this research we were able to develop the efficient, though quite simple method of atherosclerosis diagnostics, which doesn't assume the execution of any complex laboratory or clinical tests or hospitalizations. We also managed to identify the set of atherosclerosis predictors, highly associated with this diagnosis (heart rate, arterial hypertension, diabetes mellitus, age, height, weight, chronic heart failure). The details and results of this research can be found in [10,11]

2.2 Automated experiments based on Microsoft Azure Machine Learning platform using the MIMIC-III data

MIMIC-III sample of a patients having atherosclerosis. MIMIC-III database [16] is a large, single-center database comprising information relating to patients admitted to critical care units at a large tertiary care hospital in Boston (Massachusetts).

The multiple experiments (classification models training, testing and validation iterations) were run using the MIMIC-III dataset sample and such classification methods as RandomForest, ExtremeRandomTrees, XGBoostClassifier (the detailed description of these algorithms are provided in [11]), as well as gradient boosting framework LightGBM [17].

3 The results of atherosclerosis diagnostics task solution

The building of the atherosclerosis classification models based on the MIMIC-III dataset using the Microsoft Azure Machine Learning cloud-based platform allowed to carry out a series of iterations for building the classification models using such frameworks and methods as LightGBM, XGBoostClassifier, RandomForest, and ExtremeRandomTrees. The models were compared primarily based on the weighted AUC metric. The model with the maximum weighted AUC (0.85746) was built using the LightGBM classifier and the Max-AbsScaler standardization algorithm.

4 The task of obtaining optimal treatment strategies

The importance of timely atherosclerosis diagnostics can be approved by many medical experts. However, an equally important task is the development and prescription of a most efficient treatment strategies, which can be solved successfully using the up-to-date reinforcement learning methods [18].

The development of a reinforcement learning model is the current step of this study. The most suitable dataset for such a solution is the MIMIC-III dataset containing a history of medical treatments, procedures and prescribed drugs.

5 The development of the automated cardiologist's workplace

The development and further performance improvement of machine learning models is inseparably linked with the continuous updating of a data array with the most relevant incoming information. Usually, this process can be set up using the automated systems and applications that control the data flow and history and optimize many processes that the specialists (doctors) should perform on a daily basis.

The application is considered as a tool which helps to reduce the time to perform the doctor's auxiliary operations, related to medical data input, storage, search and analysis, whose main goal is to provide the support for the primary atherosclerosis diagnostics, as well as intended to support the automatic selection of optimal treatment policies.

6 Conclusion

As a result of the current research, a complete analysis of the atherosclerosis disease was carried out. This research can be considered as a sequence of the following stages.

1. Identification of the most significant markers and predictors of atherosclerosis and building the diagnostic models of atherosclerosis using the data set of VOCD.
2. The development of a generalized model for the atherosclerosis diagnostics using a large dataset.

3. Within the current study, a system for identifying the optimal atherosclerosis treatment strategies based on the reinforcement learning methods was modelled. Optimization and deployment of this model is planned as the next stage of the study.
4. In order to support and improve the functioning of the created models, an automated cardiologist's workplace is also being developed – which is a system that allows you to collect and save the relevant medical information: prescriptions and diagnoses made by doctors, patient visits and health condition.

References

1. Tuzcu, E.M., Kapadia, S.R., Tutar, E. et al.: High prevalence of coronary atherosclerosis in asymptomatic teenagers and young adults: evidence from intravascular ultrasound. Circulation 103(22), 2705-2710 (2001).
2. Reddy, A.V.N., Krishna, C.P., Mallick, P.K. *et al.* Analyzing MRI scans to detect glioblastoma tumor using hybrid deep belief networks. *J Big Data* **7,** 35 (2020).
3. Jaiswal, A.K., Tiwari, P., Kumar, S., Gupta, D.R., Khanna, A., & Rodrigues, J.J. (2019). Identifying pneumonia in chest X-rays: A deep learning approach. *Measurement, 145*, 511-518.
4. Ross EG, Shah NH, Dalman RL, Nead KT, Cooke JP, Leeper NJ. The use of machine learning for the identification of peripheral artery disease and future mortality risk. *J Vasc Surg*. 2016;64(5):1515-1522.e3. doi:10.1016/j.jvs.2016.04.026.
5. Duval S, Massaro JM, Jaff MR, et al. An evidence-based score to detect prevalent peripheral artery disease (PAD). *Vasc Med*. 2012;17(5):342-351. doi:10.1177/1358863X12445102.
6. Khokhlov, R.A., Gaydashev, A.E., Ostroushko N.I., Gaydashev A.E.: Multi-channel volume sphygmography in cardioangiological screening of the adult population. Rational Pharmacotherapy in Cardiology 11(4), 371-379 (2015).
7. Weinberg I., et al. The Systolic Blood Pressure Difference Between Arms and Cardiovascular Disease in the Framingham Heart Study. The American Journal of Medicine 2014:127; 209-15.
8. Clark CE, Taylor RS, Shore AC, et al. Association of a difference in systolic blood pressure between arms with vascular disease and mortality: a systematic review and meta-analysis. Lancet 2012:380 (9838):218.
9. Mitchell E. Noninvasive diagnosis of arterial disease. Up-ToDate URL: https://www.uptodate.com/contents/noninvasive-diagnosis-of-arterial-disease.
10. Lvovich Y.E., Kashirina I.L., Demchenko M.V. The Use of Machine Learning Methods to Study Markers of Atherosclerosis of the Great Arteries. Information technology 26(1), 46-55 (2020).
11. Demchenko, M.V., Kashirina, I.L.: The development of the atherosclerosis diagnostic models under conditions of unbalanced classes. J. Phys.: Conf. Ser. 1479 012026 (2020).
12. Azure Machine Learning documentation https://docs.microsoft.com/en-us/azure/machine-learning/, last accessed 2020/5/31.
13. Komorowski M, Celi LA, Badawi O, Gordon AC, Faisal AA. The Artificial Intelligence Clinician learns optimal treatment strategies for sepsis in intensive care. *Nat Med*. 2018;24(11):1716-1720. doi:10.1038/s41591-018-0213-5.
14. Istepanian RSH, Al-Anzi T. m-Health 2.0: New perspectives on mobile health, machine learning and big data analytics. *Methods*. 2018;151:34-40. doi:10.1016/j.ymeth.2018.05.015.

15. Khokhlov, R.A., Gaydashev, A.E., Akhmedzhanov N.M.: Predictors of atherosclerotic lesions of limb arteries according to cardioangiological screening of the adult population. Rational Pharmacotherapy in Cardiology 11(5), 470-476 (2015).
16. Johnson, A. E. W. et al. MIMIC-III, a freely accessible critical care database. Sci. Data 3, 160035 (2016).
17. LightGBM documentation, https://lightgbm.readthedocs.io/en/latest/, last accessed 2020/5/31.
18. Sutton, R., Barto, A.: Reinforcement learning. 2nd edn. The MIT Press, Cambridge (2018).

DATA ANALYSIS IN ASTRONOMY AND SPECTRAL DATA

Data for binary stars from Gaia DR2 (Extended Abstract)

Dana Kovaleva[1], Oleg Malkov[1], Sergei Sapozhnikov[1], Dmitry Chulkov[1], and Nikolay Skvortsov[2]

[1] Institute of Astronomy, Moscow 119017, Russia,
dana@inasan.ru
[2] Institute of Informatics Problems, Federal Research Center "Computer Science and Control" of the Russian Academy of Sciences, Moscow 119333, Russia

Binary and multiple stars are the significant part of the stellar population of the Milky Way (see [1]). Various methods are used to discover and observe binaries [5]. Whether certain binary star may be observed by a certain method of observation, depends on its characteristics, both astrophysical (e.g. period, semimajor axis, masses or evolutionary stage of the components) and geometrical (distance, inclination of the orbit) (see discussion in [8]). The recovery of the unbiased sample of binaries is important to understand star formation process. Important steps regarding this problem were done in the recent decade with the use of results of dedicated surveys of binaries of various observational types. The homogeneous parallaxes provided by Hipparcos astrometric space mission [9] were used to choose the targets of many surveys listed above. However, these surveys are small, limited to hundreds of objects.

At the same time, there is more data on binary and multiple stars available in catalogues and databases related to certain methods of observations. This data is non-homogeneous and difficult for data mining. Data in these datasets is often related to different categories of objects (components, pairs or systems), cross identification for the objects is often non-existent or unreliable. The Binary and multiple star DataBase (BDB) http://bdb.inasan.ru ([3], [6]) uses the Identification List of Binaries, ILB ([4], [7]), as a master catalogue providing cross-identification for all entities of binary and multiple stars. The purpose of BDB is to enable data mining for binaries from every catalogue or database.

Second data release of Gaia space mission [2] provided large amount of homogenous, high-accuracy astrometric and photometric data. In spite of the fact that Gaia DR2 treats all stars as singles in respect to astrometric solution, it still can be used to enrich our knowledge of binary population in several ways. Many stars known to be components of binaries have received new astrometric parameters, some of these binaries are resolved [10]. Tens of thousands of prospective binaries were discovered by different authors as co-moving stars situated in space close to each other. Some of these new systems overlap with known ones.

The purpose of the current research is to compare information on binary and multiple stars before and after Gaia DR2, to identify objects in the Gaia DR2 catalogue related to binary systems, and to incorporate this data into ILB catalogue and BDB database. We also discuss how binary stars are represented in the Gaia DR2.

We have considered how data for known binary stars assembled in the ILB catalogue is represented in the Gaia DR2 source catalogue and made cross-identifications when possible. We have cross-matched the datasets of co-moving candidates for binaries discovered in Gaia DR2 with each other and ILB, compiled the new updated version of ILB enhanced with new data and new records from Gaia DR2.

It was investigated how markers of quality of astrometric and photometric solution behave for the data on known binary stars in Gaia DR2. Resolved components of binary stars exhibiting visible orbital motion within about 50 pc, and unresolved binaries at distances to 200-300 pc fail quality cuts recommended by Gaia-ESA in $45\% - 80\%$ of cases.

References

1. Duchêne, G., Kraus, A.: Stellar Multiplicity. Ann. Rev. Astron. Astrophys **51**, 269–310 (Aug 2013). https://doi.org/10.1146/annurev-astro-081710-102602
2. Gaia Collaboration, Brown, A.G.A., Vallenari, A., Prusti, T., et al.: Gaia Data Release 2. Summary of the contents and survey properties. Astron. Astrophys. **616**, A1 (Aug 2018). https://doi.org/10.1051/0004-6361/201833051
3. Kovaleva, D., Kaygorodov, P., Malkov, O., Debray, B., Oblak, E.: Binary star DataBase BDB development: structure, algorithms, and VO standards implementation. Astronomy and Computing **11**, 119–127 (Sep 2015). https://doi.org/10.1016/j.ascom.2015.02.007
4. Malkov, O., Karchevsky, A., Kaygorodov, P., Kovaleva, D.: Identification list of binaries. Baltic Astronomy **25**, 49–52 (2016)
5. Malkov, O., Kovaleva, D., Kaygorodov, P.: Observational Types of Binaries in the Binary Star Database. In: Balega, Y.Y., Kudryavtsev, D.O., Romanyuk, I.I., Yakunin, I.A. (eds.) Stars: From Collapse to Collapse. Astronomical Society of the Pacific Conference Series, vol. 510, p. 360 (Jun 2017)
6. Malkov, O., Karchevsky, A., Kaygorodov, P., Kovaleva, D., Skvortsov, N.: Binary Star Database (BDB): New Developments and Applications. Data **3**(4) (2018). https://doi.org/10.3390/data3040039, https://www.mdpi.com/2306-5729/3/4/39
7. Malkov, O., Karchevsky, A., Kovaleva, D., Kaygorodov, P., Skvortsov, N.: Catalogue of identifications of objects in binary and multiple stars. Astronomichesky Zhurnal Supplement Series (in Russian) **95**(7), 3–16 (2018). https://doi.org/10.1134/s0004629918070058, https://doi.org/10.1134/s0004629918070058
8. Tokovinin, A.: From Binaries to Multiples. I. Data on F and G Dwarfs within 67 pc of the Sun. Astron J. **147**, 86 (Apr 2014). https://doi.org/10.1088/0004-6256/147/4/86
9. van Leeuwen, F.: Validation of the new Hipparcos reduction. Astron. Astrophys. **474**(2), 653–664 (Nov 2007). https://doi.org/10.1051/0004-6361:20078357
10. Ziegler, C., Law, N.M., Baranec, C., Morton, T., Riddle, R., De Lee, N., Huber, D., Mahadevan, S., Pepper, J.: Measuring the Recoverability of Close Binaries in Gaia DR2 with the Robo-AO Kepler Survey. Astron. J. **156**(6), 259 (Dec 2018). https://doi.org/10.3847/1538-3881/aad80a

Classification problem and parameter estimating of gamma-ray bursts (Extended abstract)

Pavel Minaev[1,2][0000−0002−7437−2064] and Alexei Pozanenko[1,2,3]

[1] Space Research Institute (IKI), 84/32 Profsoyuznaya Str, 117997 Moscow, Russia
minaevp@mail.ru
[2] Moscow Institute of Physics and Technology (MIPT), Institutskiy Pereulok, 9, 141701 Dolgoprudny, Russia
[3] National Research University Higher School of Economics, 101000 Moscow, Russia

There are at least two distinct classes of Gamma-Ray Bursts (GRB) according to their progenitors: short duration and long duration bursts [5, 7, 9]. It was shown that short bursts result from compact binary merging [10–12], while long bursts are associated with core collapse supernova [13]. However, one could suspect the existence of more classes and subclasses [8]. For example, compact binary can be double neutron stars, or neutron star and black hole, which might generate gamma-ray bursts with different properties.

From another hand, gamma-ray transients are known to be produced by magnetars in Galaxy, named Soft Gamma-Repeaters (SGR) [3]. A Giant Flare from SGR can be detected from a nearby galaxy, and it can mimic for a short GRB[4]. So the classification problem is very important for correct investigation of different transient progenitors.

Gamma-ray transients are characterized by a number of parameters and known phenomenology correlations between them, obtained for well classified ones. Using these correlations, which could be unique for different classes of gamma-ray transients, we can classify an event and determine the type of its progenitor, using only temporal and spectral characteristics. We suggest the statistical classification method, based on the cluster analysis of the $E_{p,i} - E_{iso}$ correlation (the position of the maximum in the energy spectrum νF_ν in the source frame depending on the isotropic equivalent of the total energy, emitted in gamma rays, Fig. 1) and the $T_{90,i} - EH$ diagram (duration in the source frame depending on the combination of parameters $EH = E_{p,i,2}\ E_{iso,51}^{-0.4}$, Fig. 2) for the trained dataset of 323 events from [6, 1, 2, 4].

Using the known dependencies, one can not only classify the types of gamma-ray bursts, but also discriminate events that are not associated with gamma-ray bursts, but have a different physical nature. We show that GRB 200415A, originally classified as a short GRB, probably does not belong to the class of short GRBs, but it is most likely associated with the giant flare of SGR. On the other hand, we can estimate one of the unknown parameters if we assume the certain classification of the event. As an example, an estimate the redshift of the GRB 200422A source is given. We also discuss that in some cases it is possible

to give a probabilistic estimate of the unknown parameters of the source. The method could be applied to any other analogous classification problems.

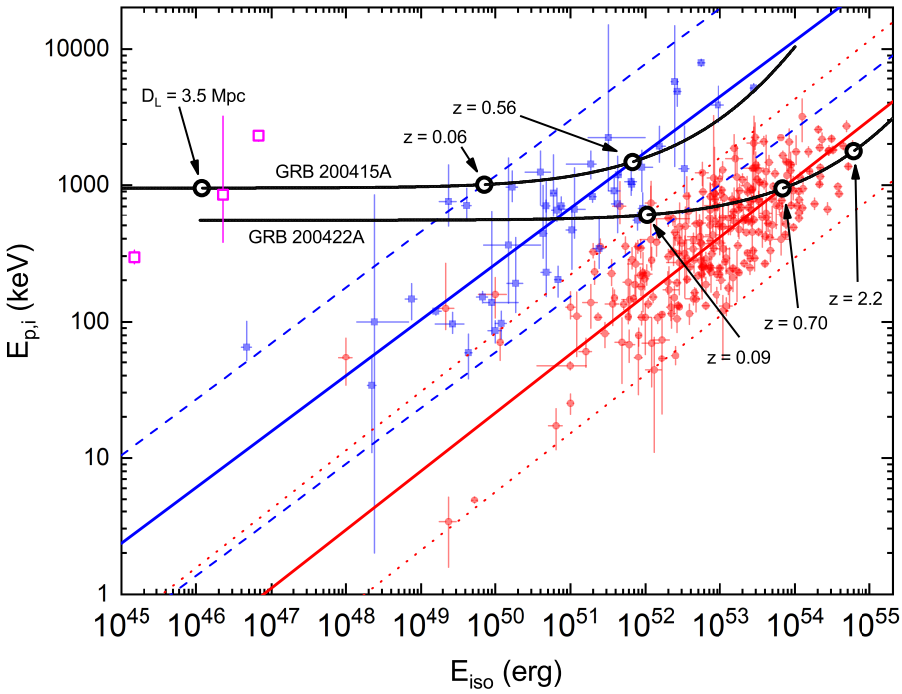


Fig. 1. The $E_{p,i}$ – E_{iso} correlation for type I bursts (blue squares), type II bursts (red circles) and SGR giant flares (magenta open squares) with corresponding fits. $2\sigma_{cor}$ correlation regions are also presented for type I (type II) bursts by blue dashed (red dotted) lines. Trajectories (dependencies on redshift) for GRB 200415A and GRB 200422A are shown by black lines with key points indicated by open circles.

References

1. Frederiks, D.D., Golenetskii, S.V., Palshin, V.D., Aptekar, R.L., Ilyinskii, V.N., Oleinik, F.P., Mazets, E.P., Cline, T.L.: Giant flare in SGR 1806-20 and its Compton reflection from the Moon. Astronomy Letters **33**(1), 1–18 (Jan 2007). https://doi.org/10.1134/S106377370701001X
2. Frederiks, D.D., Palshin, V.D., Aptekar, R.L., Golenetskii, S.V., Cline, T.L., Mazets, E.P.: On the possibility of identifying the short hard burst GRB 051103 with a giant flare from a soft gamma repeater in the M81 group of galaxies. Astronomy Letters **33**(1), 19–24 (Jan 2007). https://doi.org/10.1134/S1063773707010021
3. Kouveliotou, C., Strohmayer, T., Hurley, K., van Paradijs, J., Finger, M.H., Dieters, S., Woods, P., Thompson, C., Duncan, R.C.: Discovery of a Magnetar As-

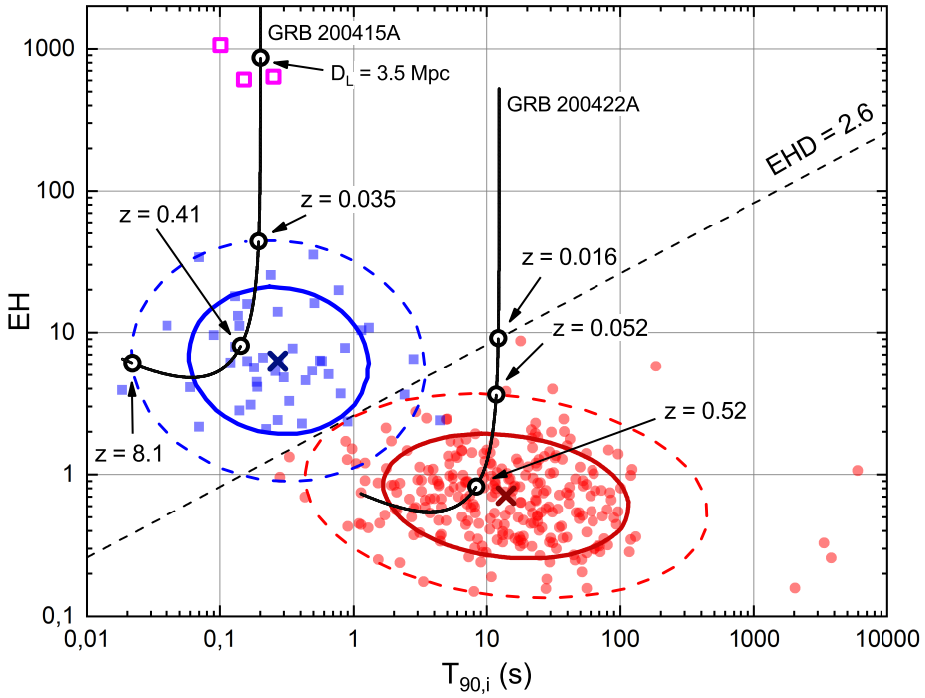


Fig. 2. The $T_{90,i}$ – EH diagram for type I bursts (blue squares), type II bursts (red circles) and SGR giant flares (magenta open squares). The dashed line ($EHD = 2.6$) is used in a blind classification of GRBs. Gaussian mixture model fits for type I and type II bursts are shown by coloured thick solid and dashed ellipses, representing 1σ and 2σ confidence regions, correspondingly. Trajectories (dependencies on redshift) for GRB 200415A and GRB 200422A are shown by black lines with key points indicated by open circles.

sociated with the Soft Gamma Repeater SGR 1900+14. ApJ **510**(2), L115–L118 (Jan 1999). https://doi.org/10.1086/311813
4. Mazets, E.P., Aptekar, R.L., Cline, T.L., Frederiks, D.D., Goldsten, J.O., Golenetskii, S.V., Hurley, K., von Kienlin, A., Pal'shin, V.D.: A Giant Flare from a Soft Gamma Repeater in the Andromeda Galaxy (M31). ApJ **680**(1), 545–549 (Jun 2008). https://doi.org/10.1086/587955
5. Mazets, E.P., Golenetskii, S.V., Ilinskii, V.N., Panov, V.N., Aptekar, R.L., Gurian, I.A., Proskura, M.P., Sokolov, I.A., Sokolova, Z.I., Kharitonova, T.V.: Catalog of cosmic gamma-ray bursts from the KONUS experiment data. I. Ap&SS **80**, 3–83 (Nov 1981). https://doi.org/10.1007/BF00649140
6. Minaev, P.Y., Pozanenko, A.S.: The $E_{p,i}$-E_{iso} correlation: type I gamma-ray bursts and the new classification method. MNRAS **492**(2), 1919–1936 (Feb 2020). https://doi.org/10.1093/mnras/stz3611
7. Minaev, P.Y., Pozanenko, A.S., Loznikov, V.M.: Extended emission from short gamma-ray bursts detected with SPI-ACS/INTEGRAL. Astronomy Letters **36**, 707–720 (Oct 2010). https://doi.org/10.1134/S1063773710100026
8. Minaev, P.Y., Pozanenko, A.S., Loznikov, V.M.: Short gamma-ray bursts in the SPI-ACS INTEGRAL experiment. Astrophysical Bulletin **65**, 326–333 (Oct 2010). https://doi.org/10.1134/S1990341310040024
9. Minaev, P.Y., Pozanenko, A.S., Molkov, S.V., Grebenev, S.A.: Catalog of short gamma-ray transients detected in the SPI/INTEGRAL experiment. Astronomy Letters **40**, 235–267 (May 2014). https://doi.org/10.1134/S106377371405003X
10. Paczynski, B.: Gamma-ray bursters at cosmological distances. ApJ **308**, L43–L46 (Sep 1986). https://doi.org/10.1086/184740
11. Pozanenko, A.S., Barkov, M.V., Minaev, P.Y., Volnova, A.A., Mazaeva, E.D., Moskvitin, A.S., Krugov, M.A., Samodurov, V.A., Loznikov, V.M., Lyutikov, M.: GRB 170817A Associated with GW170817: Multi-frequency Observations and Modeling of Prompt Gamma-Ray Emission. ApJ **852**, L30 (Jan 2018). https://doi.org/10.3847/2041-8213/aaa2f6
12. Pozanenko, A.S., Minaev, P.Y., Grebenev, S.A., Chelovekov, I.V.: Observation of the Second LIGO/Virgo Event Connected with a Binary Neutron Star Merger S190425z in the Gamma-Ray Range. Astronomy Letters **45**(11), 710–727 (Feb 2020). https://doi.org/10.1134/S1063773719110057
13. Woosley, S.E.: Gamma-ray bursts from stellar mass accretion disks around black holes. ApJ **405**, 273–277 (Mar 1993). https://doi.org/10.1086/172359

Data Quality Assessments in Large Spectral Data Collections (Extended Abstract)

A. Yu. Akhlestin, N. A. Lavrentiev, N. N. Lavrentieva, A. V. Kozodoev, E. M. Kozodoeva, A. I. Privezentsev, A. Z. Fazliev [0000-0003-2625-3156]

V. E. Zuev Institute of Atmospheric Optics SB RAS, Tomsk, Russia
lexa@iao.ru, lnick@iao.ru, lnn@iao.ru, kav@iao.ru, faz@iao.ru, remake@iao.ru, klen@iao.ru

1 Introduction

Assessment of the quality of information resources located on the Internet and representation of these resources in the form of software agents convenient for work are among currently important problems. Semantic Web (SW) technologies [1] were suggested for their solution. The tools for representing semantic annotations of resources turned out to be universal, while the techniques for assessment of the quality of resources, at least for scientific subject domains, are not universal.

Assessments of the quality of scientific resources can be divided into two groups. In the first group, the reliability of scientific resources connected with a mathematical model of a subject domain is assessed. For the validity check, criteria are used derived from the constraints imposed by the model on the entities described in the subject domain. In the second group, trust in resources is assessed, in particular, the influence of experts, presenting these resources.

When studying the states and transitions of molecules and atoms, large-volume numerical arrays are used; some of them contain parameters of several billion transitions for a molecule.

Information system W@DIS [2] is a part of VAMDC; it includes primary, expert, and empirical spectral data and provides a complete description of the results of their quality assessment. These data describe more than hundred molecules (including isotopologues). A feature of W@DIS IS is the presence of complete results of the physical parameters and data sources quality analysis for all spectral data contained in it. All results of assessment of the quality of are presented in the form of ontologies [3-5].

In this work, we discuss the sequence of application of different methods for assessment of spectral information quality, features of these methods, and user interfaces for work with the results of spectral information quality analysis.

2 Data model in quantitative spectroscopy. Data quality control

Before consideration of a data model for quantitative spectroscopy, we should emphasize the fact that the model suggested does not describe all facts related to the subject domain under study. However, it includes most data used in applied subject domains. The data model describes three data types.

Data related to measurements or calculations published in a particular work and their brief description together are called the primary data source below. Primary experimental and theoretical data are interrelated, because not all values of the properties of transitions and states are measurable, for example, quantum numbers, the values of which can be determined only by solving computational problems. In addition to primary data, composite data are also used in spectroscopy. Two composite data types are important for applications: empirical and expert data. The former are calculated by using a multiset of measured data, and the latter are a set of calculated, measured, and reference data. Thus, the data layer contains three data types: measurement data and their derivatives (empirical energy levels calculated on the basis of the Rydberg–Ritz principle [6]), calculation data, and expert data, which are a mixture of the first and second data types.

The analysis of the quality of C^0 — the multiset of transitions characteristics of which are measured and published — for each molecule takes the central place in the data analysis. This set consists of several parts. The part C^0_A includes incorrectly identified transitions which are excluded from the quality analysis. The part C consists of correctly identified transitions, but can contain conflicting values of wavenumbers and other characteristics for identical transitions. For a number of molecules, Cantor sets of states (sets with all elements unique) are composed based on the multiset of transitions measured. The states in these sets are described by empirical energy levels E_R (for example, for a water molecule), which can be used for the check of the corresponding set of transitions C. For this, all possible transitions (C_R), identical to the transitions from the set C, should be constructed from the empirical E_R levels. These transitions, which are identical to the transitions from C_R, form the set of transitions C^i_R. It consists of two subsets A^i and A_i, which include the transitions from C with the wavenumbers, differing from those of identical transitions from C_R by less or more than Δ cm^{-1}, respectively. Here Δ is the maximum permissible deviation between the wavenumbers of transitions from the set C^i_R and of identical transitions from the set C_R. Let us write the obvious equalities: $C^i_R = A^i \cup A_i$, $C = C_A \cup C^i_R$, and $C^0 = C \cup C^0_A$ (U is the union of sets).

The numbers of elements in the sets C^i_R and C_A do not change at a fixed number of the elements in the multiset C, but the number of elements in the sets A^i and A_i depends on Δ. This parameter is used as a quantitative measure of the permissible difference between the values of a wavenumber which describes a transition.

The quality analysis of data related to transitions from the set C_A is of greatest interest. If this set is Cantor, then a possibility of compiling a set of empirical energy levels from it is doubtful. Therefore, it is necessary to carry out measurements to expand the number of empirical energy levels in the set C^i_R.

The main information resources in W@DIS are sources of data on states and transitions. Each data source includes an uploaded numerical array or plot published and the properties of this array (plot).
Let D^0_n and D^0_{Em} denote the primary array of measured values of physical parameters and the expert data array, respectively. Note that the above introduced multiset of transitions is the union of arrays:

$$C^0 = U^N_{n=1} D^0_{n.} \quad (1)$$

Hence, the equalities

$$C = U^N_{n=1} D_n,\ C^0_A = U^M_{m=1} D_m, \text{ etc.} \quad (2)$$

are true.

3 General description of the collection of spectral line parameters

The spectral data quality is analyzed in each paper on quantitative spectroscopy with the use of traditional methods. The division of spectral data into data sets, which relate to individual molecules and solutions to one of the seven spectroscopy problems in the W@DIS information system (IS) (http://wadis.saga.iao.ru) allowed us to simplify the primary data analysis and to assess trust in expert data used in different application problems. Existing expert data created by different experts are contradictory. To resolve the contradictions, a decomposition method has been suggested to assess the trust in these data. The methods for the data quality analysis and assessment of trust in expert spectral data are briefly described in the report. Spectral data collections in W@DIS are associated with the output data of several dozens of atmospheric molecules. The structure of a collection is identical to the structure of physical problems solved in molecular spectroscopy and includes data on energy levels, vacuum transitions, and spectral parameters of lines of individual molecules and their mixtures. Each part of a collection is assigned to the publication from which these data has been extracted. Along with data sources, the system includes all publications related to data arrays accumulated in the collections.
Each data source is connected with a conventional set of metadata, which describes the set of properties of this source. Data on different molecules are not connected. This fact determines the branched modular structure of ontologies, each describing one of the modules. Every molecular collection is described by ontologies, which characterizes the information resources or physical quantities in this collection.
The main properties of data sources are those which describe the source quality analysis results. Data sources in the collections are typified. The following types are used for the classification: primary (measurements or calculations), expert, and empirical. The first three types are traditional for natural science data collections, while empirical data in different subject domains can be formed according to different principles. In spectroscopy, empirical data includes data related to energy levels and other physi-

cal quantities. Empirical energy levels are obtained from processing the complete set of energy levels derived from the inverse problem solution accounting the Rydberg–Ritz principle.

4 Conclusions

We describe a collection of spectral data in W@DIS information system and the methods used in it for the analysis of spectral data quality. A data model in quantitative spectroscopy and a group of applications for data acquisition are presented. The sequence of actions for the data quality analysis is explained. The quality assessment techniques commonly used in spectroscopy are briefly described, as well as three methods for the spectral data quality analysis developed by the authors for large arrays of conflicting spectral data. For better visual representation of the results, graphical user interfaces for working with the results of data quality assessment are exemplified. The percentage ratio of the parts of the multiset is shown, with the multiset of transitions of the main water molecule isotopologue as an example. In particular, we have shown that less than half of a percent of the data remains after data filtering in the automatic spectral data quality analysis, which require additional processing by experts.

The main part of the work was performed within project № AAAA-A17-117021310147-0.

References

1. Berners-Lee, T., Hendler, J., Lassila,O.: The Semantic Web. Scientific American, May 17, 2001.
2. Akhlyostin, A., Apanovich, Z., Fazliev, A., Kozodoev, A., Lavrentiev, N., Privezentsev, A., Rodimova, O., Voronina, S., Csaszar, A.G., Tennyson, J.: The current status of the W@DIS information system. In: Matvienko, G., Romanovskii, O. (eds.) Proc. SPIE of 22-nd International Symposium Atmospheric and Ocean Optics: Atmospheric Physics, 10035, 100350D (2016).
3. Privezentsev A.I., Tsarkov D.V., Fazliev A.Z.: Computed knowledge base for description of information resources of molecular spectroscopy. 3. Basic and applied ontologies. Digital Library Journal, v.15(2) (2012).
4. Voronina, S.S., Privezentsev, A.I., Tsarkov, D.V., Fazliev, A.Z.: An Ontological Description of States and Transitions in Quantitative Spectroscopy. In: Proc. of SPIE XX-th International Symposium on Atmospheric and Ocean Optics: Atmospheric Physics, v. 9292, 92920C (2014).
5. Fazliev, A., Privezentsev, A., Tsarkov, D., Tennyson, J.: Ontology-Based Content Trust Support of Expert Information Resources in Quantitative Spectroscopy. In: Knowledge Engineering and the Semantic Web, Communications in Computer and Information Science, v. 394, Springer Berlin Heidelberg, pp.15-28 (2013).
6. Watson, J.K.G.: The Use of Term-Value Fits in Testing Spectroscopic Assignments. Journal of Molecular Spectroscopy, v. 165, 283-290 (1994).

High-Dimensional Simulation Processes in New Energy Theory: Experimental Research (Extended Abstract)

Elena Smirnova [0000-0003-2275-3276], Vladimir Syuzev[0000-0002-8689-8282], Roman Samarev [0000-0001-9391-2444], Ivan Deykin[0000-0002-8151-6337] and Andrey Proletarsky[0000-0001-8060-3431]

Bauman Moscow State Technical University, Moscow 105005, Russia

The paper is devoted to the fundamental problem solution on an effective multidimensional digital signal processing development in the framework of research work supported by the Russian Federation Ministry of Science and Higher Education (Project #0705-2020-0041), related to the new materials for a new memory generation using direct recording methods with ultrashort laser pulses to solve problems provided by the end-to-end digital technology in the fields of Neurotechnology and Artificial Intelligence.

The relevance of the research topic described in this article is due to the need of new scientific methodology development for the synthesis of high-precision and high-performance algorithms for simulating deterministic and random signals of large dimensions within the framework of classical and generalized correlation theory in the spectral domain of harmonic bases, using the original algorithmic relationship of models of pseudo-random and deterministic signals, which will allow to create common simulation models on a single mathematical and software basis, as well as will provide an effective tool for statistical research of real-time systems for various purposes.

There is a need to develop a new energy theory of digital matrix representation and conversion of real signals to create fast algorithms that reduce computational complexity and improve the accuracy of signal recovery. The spectral algorithms are described for simulating multi-dimensional discrete deterministic and random bandpass signals on the example of two-dimensional discrete basis functions and Fourier and Hartley transformations, accounting the energy characteristics and autocorrelation functions of these signals. Communication equations are given that allow authors to develop two-dimensional simulation algorithms for signals with a non-axial frequency spectrum.

There are certain forms of data that are especially well studied and therefore have general methods of analysis, such as time series, working with large amounts of data [1]. In other cases, the input data is more complex in shape or size, however we can get data from any source nowadays, starting from the genome [2] and ending with the media [3], in these cases more complex data get a more specific approach.

Big data means not only extended in computational space but is also in time. Time data characteristics may render traditional algorithms fundamentally useless. New data could come continuously and it could be not only required to be stored [4], there

should be a method of dividing the input data into operational events in real time with further intent of using these events for forecasting [5]. Another level of complication is a multidimensional data. Finance uses a multidimensional dynamic analysis [6], the response timelines to threats in computer networks [7], articulated thinking visualization, all these areas need new approach of research.

One of the problems with multidimensional analysis stems from it being visually unintuitive. While it is possible to imagine the nature behind time series or to locate the connection between real life phenomenon and its two-, three- or even four-dimensional representation, higher numbers of dimensions may often lead to confusion. However, computational power available today does not only allow us to finally work with data as big as it comes but also allows us to disengage from our own biases by placing more work onto the machine. Indeed, such areas as machine learning and neural networks are not limited by executing calculations preassigned by the researcher but can to some extent choose their own mode of actions claiming levels of flexibility previously unattainable.

Modeling and simulation grant us the ability to study any real time processes virtually saving costs for the physical experiments. The usage of higher dimensions contributes to the accuracy delivered by the new algorithms. The usage of energy spectra places the scientific research for these algorithms in the well-researched area of spectral theory [7 - 10].

Spectral theory provides new approaches of research based on matrix mathematical apparatus that renders the new algorithms prepared for further automatization by the means of tool developed in such areas as machine learning, artificial intelligence, and big data processing.

The authors consider the basics of the theory of spectral simulation of multidimensional signals in harmonic bases continuing their research [11 – 15]. Properties of two-dimensional harmonic discrete basis functions and transformations necessary for the development of spectral algorithms for simulating two-dimensional signals are considered in this article as well as its experimental software realization.

Accuracy can be measured by comparing with theoretical values for a specific signal, and speed can depend on the mathematical form of the algorithm – equations are computed faster than complex algorithms using matrices. Mathematical equations provide low memory size requirements as well. A possible disadvantage of the described algorithms is the need for impressive mathematical training before programming. The output data of the software is the theoretical, algorithmic, and experimental autocorrelation functions, as well as the errors, as the difference between autocorrelation functions, and the simulated signal itself. The developed application allows to simulate signals based on specified characteristics with the ability to simulate one- and two-dimensional signals.

The software implementation has been done using Lazarus IDE, supporting Free Pascal programming language. The chosen IDE provides all the necessary mathematical functions and instruments allowing to build desired interfaces. Computational complexities for calculating algorithmic and experimental are both O(M*(N2-N1)). Thus, computational complexity reaches only O(M*(N2-N1)), which is closer to linear time complexity rather than to $O(M^2)$ – such result may be deemed sufficient. So,

this paper starts new spectra theory development for high-dimensional signal simulation. First steps using two-dimensional simulation proofed new direction of research. The paper shows that the spectral theory provides new approaches of research based on matrix mathematical apparatus. The new algorithms could be prepared for further automatization by the means of tool developed in such areas as machine learning, artificial intelligence, and big data processing.

The method of two-dimensional simulation of signals in a complex basis reduces algorithmizing to the execution of pre-derived mathematical formulas, which reduces the computational complexity and resource intensity of the algorithm, and the use of linear data structures positively affect the scalability of the developed solution. The use of a complex basis that more accurately describes the nature of ongoing processes provides higher simulation accuracy.

The software solution implemented in the Lazarus environment in the Free Pascal language meets the requirements and allows generating deterministic and random signals, as well as evaluating the quality of simulation by displaying the error graph and/or displaying the average error number.

The simulation method in the Hartley basis and the software solution based on it, as in the case of the complex basis, make the algorithm less resource intensive. The software solution is implemented in Microsoft Visual Studio using the C# language. For both bases, both deterministic and random signals can be simulated (at the user's discretion), and the shape of the SPDF signal can be selected: rectangular and rectangular-triangular. The signals obtained meet the expectations for the spectrum, and when compared with theoretical and algorithmic ACF, they show error levels of less than 0.05. In future studies, it is planned to expand the choice of signal forms, as well as to test new methods on a wider range of tasks. It is also planned to create a library of obtained algorithms combined in a single solution that provides simulation in different bases.

References

1. Kraeva Y., Zymbler M. (2019) Scalable Algorithm for Subsequence Similarity Search in Very Large Time Series Data on Cluster of Phi KNL. In: Manolopoulos Y., Stupnikov S. (eds) Data Analytics and Management in Data Intensive Domains. DAMDID/RCDL 2018. Communications in Computer and Information Science, vol 1003. Springer, Cham.
2. Ceri S. et al. (2018) Overview of GeCo: A Project for Exploring and Integrating Signals from the Genome. In: Kalinichenko L., Manolopoulos Y., Malkov O., Skvortsov N., Stupnikov S., Sukhomlin V. (eds) Data Analytics and Management in Data Intensive Domains. DAMDID/RCDL 2017. Communications in Computer and Information Science, vol 822. Springer, Cham.
3. Tikhomirov M., Dobrov B. (2018) News Timeline Generation: Accounting for Structural Aspects and Temporal Nature of News Stream. In: Kalinichenko L., Manolopoulos Y., Malkov O., Skvortsov N., Stupnikov S., Sukhomlin V. (eds) Data Analytics and Management in Data Intensive Domains. DAMDID/RCDL 2017. Communications in Computer and Information Science, vol 822. Springer, Cham
4. Chernenkiy V.M. et al. (2019) The Principles and the Conceptual Architecture of the Metagraph Storage System. In: Manolopoulos Y., Stupnikov S. (eds) Data Analytics and

Management in Data Intensive Domains. DAMDID/RCDL 2018. Communications in Computer and Information Science, vol 1003. Springer, Cham.
5. Andreev A., Berezkin D., Kozlov I. (2018) Approach to Forecasting the Development of Situations Based on Event Detection in Heterogeneous Data Streams. In: Kalinichenko L., Manolopoulos Y., Malkov O., Skvortsov N., Stupnikov S., Sukhomlin V. (eds) Data Analytics and Management in Data Intensive Domains. DAMDID/RCDL 2017. Communications in Computer and Information Science, vol 822. Springer, Cham.
6. Popov D.D., Milman I.E., Pilyugin V.V., Pasko A.A. (2017) Visual Analytics of Multidimensional Dynamic Data with a Financial Case Study. In: Kalinichenko L., Kuznetsov S., Manolopoulos Y. (eds) Data Analytics and Management in Data Intensive Domains. DAMDID/RCDL 2016. Communications in Computer and Information Science, vol 706. Springer, Cham.
7. Skryl', S., Sychev, M., Sychev, A., Mescheryakova, T., Ushakova, A., Abacharaeva, E., Smirnova, E. (2019) Assessing the Response Timeliness to Threats as an Important Element of Cybersecurity: Theoretical Foundations and Research Model / Communications in Computer and Information Science, Vol. 1084, 2019, pp. 258-269 //3rd Conference on Creativity in Intelligent Technologies and Data Science, CIT and DS 2019; Volgograd; Russian Federation; 16 September 2019 till 19 September 2019.
8. Syuzev V., Smirnova E., Gurenko V. Speed Algorithms for Signal's Simulation / Science Problems. 2018. №11 (131). URL: https://cyberleninka.ru/article/n/bystrye-algoritmy-modelirovaniya-signalov (Date of retrieve: 04.05.2020) - in Rus.
9. Syuzev V.V., Gurenko V.V., Smirnova E.V. Signal Simulation Spectra Algorithms as Learning and Methodical Tools of Engineers' Preparation // Machinery and Computer Technologies. 2016. №7. URL: https://cyberleninka.ru/article/n/spektralnye-algoritmy-imitatsii-signalov-kak-uchebno-metodicheskiy-instrument-podgotovki-inzhenerov (Date of retrieve: 04.05.2020) - in Rus.
10. Rusnachenko N., Loukachevitch N. (2019) Neural Network Approach for Extracting Aggregated Opinions from Analytical Articles. In: Manolopoulos Y., Stupnikov S. (eds) Data Analytics and Management in Data Intensive Domains. DAMDID/RCDL 2018. Communications in Computer and Information Science, vol 1003. Springer, Cham. https://link.springer.com/chapter/10.1007/978-3-030-23584-0_10
11. A. Sotnikov, A. Proletarsky, V. Suzev, T. Kim (2019) APPLICATION OF REAL-TIME SIMULATION SYSTEM IN THE FRAMEWORK OF DIGITAL SIGNAL PROCESSING PRACTICECOURSE, EDULEARN19 Proceedings, pp. 5812-5820.
12. Syuzev, V., Smirnova, E., Kucherov, K., Gurenko, V., Khachatryan, G. Spectral Algorithms for Signal Generation as Learning-Methodical Tool in Engineers Preparation / In Smirnova, E. and Clark, R. (ed) Analyzing Engineering Education in a Global Context, IGI Global, 38, 2019. Pp. 254-263. DOI: 10.4018/978-1-5225-3395-5.
13. Kim T., Smirnova E., Sotnikov A., Syuzev V. (2019) Articulated Thinking as the Mean of the Digital Signal Processing Methods Research, INTED2019 Proceedings, pp. 5241-5246.
14. Smirnova E., Syuzev V., Gurenko V., Bychkov B. (2019) Spectral Signal Simulation as a Scientific and Practical Task in the Training of Engineers, INTED2019 Proceedings, pp. 4511-4516.
15. Deykin I., Syuzev V., Gurenko V., Smirnova E., Lubavsky A. (2019) Random Bandpass Signals Simulation with Complex Basis Algorithm / Science Problems Journal. 2019. № 11 (144) – in Rus.

Databases of Gamma-Ray Bursts' Optical Observations (Extended Abstract)

Alina Volnova[1], Alexei Pozanenko[1,2,3], Elena Mazaeva[1], Sergei Belkin[1], Namkhay Tungalag[4], and Pavel Minaev[1]

[1] Space Research Institute of the Russian Academy of Sciences (IKI), 84/32 Profsoyuznaya Street, Moscow, 117997, Russia

[2] National Research University Higher School of Economics, Myasnitskaya 20, 101000, Moscow, Russia

[3] Moscow Institute of Physics and Technology (MIPT), Institutskiy Pereulok, 9, Dolgoprudny, 141701, Russia

[4] Research Center of Astronomy & Geophysics, of the Mongolian Academy of Sciences, P.O.Box-152, Ulaanbaatar, 14200, Mongolia

1 Introduction

Gamma-ray bursts (GRBs) are the most violent extragalactic explosions in the Universe, releasing 10^{48}–10^{54} ergs in gamma rays typically within a few seconds, and up to a few hours in some instances [1]. GRBs may be divided into two classes regarding their nature: GRBs of Type I, with duration less than 2 seconds, harder spectrum, and caused by merging of close binary systems with at least one neutron star; and GRBs of Type II, with duration longer than 2 seconds, softer spectrum, and caused by death of a very massive star during core-collapse (e.g. [2,3]). An internal dissipation process within the jet is thought to produce prompt gamma-ray emission [4], while a longer lived, multiwavelength afterglow is expected to be produced as the jet propagates through the circumstellar medium (of constant density or a stellar-wind-like density [5]). In 2/3 of cases, GRBs are accompanied with an X-ray afterglow and in about 40% of cases an optical afterglow is discovered (e.g. [6]). Nearby long GRBs exhibit the feature of the supernova (SN) of Type Ic in the light curve which confirms the core-collapse nature of these events [7]. Short GRBs show the kilonova feature originating from the merging compact binary system of binary neutron star (e.g. [8,9]). Both kilonovae and supernovae allow astronomers to investigate the processes of nucleosynthesis and building matter in the Universe. When the optical afterglow fades out, the host galaxy of the burst may be observed, which helps investigating the environment of GRBs, and may be the only way to estimate the distance to the event [10].

Optical data are crucial for GRB investigations, because it allows learning physical properties of the circumburst medium and the burst ejecta. Statistical studies of optical properties of a big number of bursts may give constrains on the evolution of the Universe. Since the discovery of the first optical counterpart in 1997, GRBs were observed by many different instruments in all ranges of the spectrum, and more than 1300 X-ray and about 800 optical afterglows were discovered [6].

In this work, we discuss databases of raw optical observations of the GRBs and databases containing results of statistical studies of their properties, obtained mostly with optical data. We present the unique database of our Space Research Institute Gamma-Ray Burst Follow-up Network (IKI GRB-FuN) starting from 2001 and containing optical data for more than 500 GRBs.

2 IKI GRB-FuN observations and data collection (database)

Statistical analysis along with investigations of very bright GRB is the main tools of GRB investigation. Bright GRB can support us by a dense photometric multicolor light curve, spectroscopy, and polarimetry. However, bright bursts in optic can be counted on the fingers of one hand. Statistical analysis of many combined light curves until now is the main instrument of investigation of a «poor man», and the distributions of the main light curve parameters allow estimating the GRB physical properties according to suggested models. To have a robust data for statistical investigation one need to have a verified data collections / databases.

General optical light curves of GRB consist of several common elements originated from the physics of the phenomenon. Statistical dependences inherent in the prompt emission phase are linked to the processes of the burst itself, so it helps to study the burst energetics [11,12]. Many investigations to search for and to study the statistical properties of afterglows and jet breaks have been carried out based the optical data [13]. SNe related to the GRBs are also collected to study their statistical properties [14]. Statistical properties of short GRBs are collected in a few papers [2,15,16]. Until now, no dependence is found between prompt gamma-ray emission parameters and optical properties of GRB. Therefore, special interest suggests databases comprising both gamma-ray and optic data [17].

The Space Research Institute Gamma-Ray Burst Follow-up Network (IKI GRB-FuN) started operation in 2001. The main idea of the network is using dedicated time on the existed facilities, i.e. IKI GRB-FuN is overlay network. A core of the network in Space Research Institute (IKI) is automatically sending target-of-opportunity applications to different observatories, planning regular observation of afterglow, searching for SN featuring photometric light curve and in spectrum, searching and observing host galaxies. The most prominent result of the IKI GRB-FuN is the observation of GRB 170817 related to the first detection of gravitational-wave event GW170817 of binary neutron star merging [8,9]. Nowadays the network comprises of about 25 telescopes with aperture from 0.2 to 2.6 meters located in different observatories all over the world; the IKI GRB-FuN is also collaborating with ISON network [18] and other observatories by submitting proposals for large aperture telescopes (Crimea, Mondy, Kislovodsk, N.Arkhyz, Terskol – Russia; Kazakhstan; Uzbekistan; Armenia; Georgia; Mongolia; Ovalle, Chile; Sutherland, South Africa; Nainital, India; Australia; Canary Islands; Hawaii). The distribution of the observatories with the longitude allows observing the optical afterglow of the GRB almost without interruption throughout the day and building detailed light curves. All the received data are collected in the Space Research Institute in Moscow.

The database of IKI GRB-FuN comprise multicolor image observations of different optical transients: Gamma-ray bursts optical counterparts, Tidal Disruption Event (TDE), Supernova (SN), Soft Gamma-Rays repeaters (SGR), unclassified transients and observations of the region localization of Gravitational Wave (GW) events registering by LIGO/Virgo in O2 and O3 runs [19]. Database counts more than 500 GRBs with at least one observation available, and 20% of the objects have light curves with more than 10 photometry data. The observations obtained mostly by telescope/observatories that are given in Table 1 and include observations of GRBs in different phases: search for optical counterpart, a few prompt observations, early and late time afterglow observations, supernovae (13) and candidate in supernovae associated with GRBs (4), kilonovae (3), and host galaxies of GRBs (48). Database contains raw data of series of the object observations, calibration data and a calibrated stack image of time series per each epoch. Besides of raw data database contains preliminary photometry of object or an upper limit at the place of supposed object. Many observations of GRB results in upper limit only, which ranges from 16m up to 24m in R-filter. The most of observations performed in R-filter, which is accepted de-facto for observations of GRB afterglow.

Using the same filter permits to construct uniform light curve if observations produced by different telescopes/observatories. It is essential to use the same calibration stars for differential photometry for all observations of an object in different telescopes. Calibration stars in the field of view is determined by automatic procedure, which we developed for uniform photometric calibration [20]. The calibration stars are also included in our database. The access to the database may be arranged via FTP protocol, and a specific username and password should be requested by the e-mail to the database PI Alexei Pozanenko (apozanen@iki.rssi.ru).

3 Discussion

In discussion of a database of GRB in optic, we would like to follow criterion suggested in [21,22]. The most sensitive issues are open access, first-hand presentation of data by the experimenter (i.e., original), and uniformity. The two first statements are self explained. The uniformity of the data relies heavily on the use of the same instrument to obtain experimental data. In addition, completeness is important. The notion of completeness of sampling implies the uniformity of data to a certain limit of sensitivity, and the estimating of possible selective effects that impede this uniformity.

There are different phases of GRB emission. Prompt phase should be recorded with fine time sampling comparable with time scale of elementary structures (pulses) of gamma-ray observed in gamma-rays. These are the most desirable, but also the most technically sophisticated observations. There are no such observations yet.

The phase of SN associated with GRB is still not well represented. SN can be effectively detected up to a distance equivalent to a redshift of $z \sim 0.5$. However, the number of GRBs up to the distance is small, about 2 per year (cf. 100 GRBs per year registered in gamma-rays). The multicolor photometry database for afterglow and SN are still not developed.

Despite on the GRB initially is detected by gamma-ray telescopes, emission from a source of GRB registered virtually by in every energy wavelength, from radio- to TeV emission. In a few cases, detected electromagnetic emission spans over 14 orders. It is obvious that synchronous observations in radio, infrared, optic, X-ray and gamma-rays give a lot of information for modeling physical processes of gamma-ray bursts. However, till now there is no a comprehensive database comprising all wavelengths. Compiling such a database is a matter of the near future.

One of the most interesting problems of transient astronomy is observations of GRBs related to gravitational wave events discovered by LIGO/Virgo/KAGRA. Short duration GRB is expected after binary neutron star merging (BNS). The first BNS registered by LIGO/Virgo [23] occurred to be extremely lucky by detection of an electromagnetic counterpart [8]. In particular, in optic the electromagnetic counterpart was detected as a kilonova. The second one BNS GW190425 detected by LIGO/Virgo [24] was less successful and could be detected only as GRB 190425 in gamma rays. [25]. In this case, the source localization area was huge. Several optical surveys have been carried out covering a significant part of the localization region (but not all), as well as target galaxies in the expected volume of localization. IKI GRB FuN is also participated in the search. No optical counterpart was found in any survey. In any case, one needs to construct database of each survey, which includes details of target and mosaic observation of particular GW error localization area. The task of finding the optical counterpart in huge areas of localization has proved to be more difficult than finding a needle in a haystack.

References

1. Greiner J., Mazzali P. A., Kann D. A., et al. A very luminous magnetar-powered supernova associated with an ultra-long γ-ray burst. // Nature, 523, 189 (2015).
2. Kann D. A., Klose S., Zhang B., et al. The Afterglows of Swift-era Gamma-Ray Bursts. II. Type I GRB versus Type II GRB Optical Afterglows. // Astrophys. J., 734, 96 (2011).
3. Kumar P. & Zhang B. The physics of gamma-ray bursts & relativistic jets. // Phys. Rep., 561, 1 (2015).
4. Hu Y.-D., Liang E.-W., Xi S.-Q., et al. Internal Energy Dissipation of Gamma-Ray Bursts Observed with Swift: Precursors, Prompt Gamma-Rays, Extended Emission, and Late X-Ray Flares. // Astrophys. J., 789, 145 (2014).
5. Mészáros P., & Rees M. J. Optical and Long-Wavelength Afterglow from Gamma-Ray Bursts. // Astrophys. J., 476, 232 (1997).
6. J. Greiner's GRB webpage; http://www.mpe.mpg.de/~jcg/grbgen.html
7. Hjorth J. & Bloom J. S. The Gamma-Ray Burst - Supernova Connection. // Chapter 9 in "Gamma-Ray Bursts", Cambridge Astrophysics Series 51, eds. C. Kouveliotou, R. A. M. J. Wijers and S. Woosley, Cambridge University Press (Cambridge), p. 169-190 (2012).
8. Abbott B.P., Abbott R., Abbott T.D., et al. Multi-messenger Observations of a Binary Neutron Star Merger. // Astrophys. J. Lett. 848, L12 (2017).
9. Pozanenko A. S., Barkov M. V., Minaev P. Yu., et al. GRB 170817A Associated with GW170817: Multi-frequency Observations and Modeling of Prompt Gamma-Ray Emission. // Astrophys. J. Lett., 852, L30 (2018).

10. Volnova, A. A., Pozanenko, A. S. Gorosabel, J., et al. GRB 051008: a long, spectrally hard dust-obscured GRB in a Lyman-break galaxy at z ≈ 2.8. // Mon. Not. Roy. Astron. Soc., 442, 2586 (2014).
11. Amati L., Guidorzi C., Frontera F., et al. Measuring the cosmological parameters with the $E_{p,i}$-E_{iso} correlation of gamma-ray bursts. // Mon. Not. Roy. Astron. Soc., 391, 577 (2008).
12. Ghirlanda G., Nava L., Ghisellini G., et al. Gamma-ray bursts in the comoving frame. // MNRAS, 420, 483 (2012).
13. Wang X.-G., Zhang B., Liang E.-W., et al. Gamma-Ray Burst Jet Breaks Revisited. // Astrophys. J., 859, 160 (2018).
14. Cano Z. The Observer's Guide to the Gamma-Ray Burst-Supernova Connection. // Eighth Huntsville Gamma-Ray Burst Symposium, held 24-28 October 2016 in Huntsville, Alabama. LPI Contribution No. 1962, id.4116 (2016).
15. Pandey S. B., Hu Y., Castro-Tirado A. J., et al. A multiwavelength analysis of a collection of short-duration GRBs observed between 2012 and 2015. // Mon. Not. Roy. Astron. Soc., 485, 5294 (2019).
16. Minaev P. Y., Pozanenko A. S. The $E_{p,I}$-E_{iso} correlation: type I gamma-ray bursts and the new classification method. // Mon. Not. Roy. Astron. Soc., 492, 1919 (2020).
17. Wang F., Zou Y.-C., Liu F., et al. A Comprehensive Statistical Study of Gamma-Ray Bursts. // Astrophys. J., 893, id.77 (2020).
18. Pozanenko A., Mazaeva E., Volnova A., et al. GRB Afterglow Observations by International Scientific Optical Network (ISON). // Eighth Huntsville Gamma-Ray Burst Symposium, held 24-28 October, 2016 in Huntsville, Alabama. LPI Contribution No. 1962, id.4074 (2016).
19. Mazaeva E., Pozanenko A., Volnova A., et al. Searching for Optical Counterparts of LIGO/Virgo Events in O2 Run. // Communications in Computer and Information Science, 1223, 124 (2020).
20. Skvortsov N. A., Avvakumova E. A., Bryukhov D. O., et al. Conceptual approach to astronomical problems. // Astrophys. Bull., 71, 114 (2016).
21. Kalinichenko L. A., Volnova A. A., Gordov E. P., et al. Data access challenges for data intensive research in Russia. // Informatics and Applications, 10, 2 (2016).
22. Kalinichenko L., Fazliev A., Gordov E. P., et al New Data Access Challenges for Data Intensive Research in Russia. // CEUR Workshop Proceedings "Selected Papers of the 17th International Conference on Data Analytics and Management in Data Intensive Domains, DAMDID/RCDL 2015", p. 215 (2015).
23. Abbott, B.P., Abbott, R., Abbott, T.D. et al. Gravitational Waves and Gamma-Rays from a Binary Neutron Star Merger: GW170817 and GRB 170817A. // Astrophysical Journal Letters 848, L13 (2017).
24. Abbott, B.P., Abbott, R., Abbott, T.D. et al. GW190425: Observation of a Compact Binary Coalescence with Total Mass ∼ 3.4 $M_{\odot}$. // Astrophysical Journal Letters, 892, id.L3 (2020).
25. Pozanenko A. S., Minaev P. Yu., Grebenev S. A. Chelovekov I. V. Observation of the Second LIGO/Virgo Event Connected with a Binary Neutron Star Merger S190425z in the Gamma-Ray Range. // Astron. Lett., 45, 710 (2020).

Variable stars classification with the help of Machine Learning (Extended Abstract)

K. Naydenkin[1], K. Malanchev[2,3], and M. Pruzhinskaya[3]

[1] Physics Faculty, Lomonosov Moscow State University, Leninskii Gori 1, 119234 Russia, rtut654@gmail.com
[2] Lomonosov Moscow State University, Sternberg Astronomical Institute, Universitetsky pr. 13, Moscow, 119234, Russia
[3] National Research University Higher School of Economics, 21/4 Staraya Basmannaya Ulitsa, Moscow, 105066, Russia

With the appearance of modern technologies such as CCD-matrices, large telescopes and computer networks the precision of our observations increased immensely. On the other hand, such accurate and complex data formed TBs large data bases which are very fragile and unattainable for the treatment by classical methods. The scales of this problem can be seen especially in variable star sky surveys. For many terabytes of data one has to classify all the stars in catalog to find stars of particular type of variability. This problem is known as very important since almost every part of modern astrophysics is interested in new objects to study. In some fields like cosmology, this question is very vital due to high demand for additional data of model-anchors like Cepheids or supernovae stars. To facilitate this task many Machine Learning based algorithms were proposed (Richards et al., 2011 [2]). In this study we perform a way to classify the Zwicky Transient Facility Public Data Release 1 catalog onto variable stars of different types. As the priority classes we set Cepheids, RR Lyrae and $\delta Scuti$.

The Zwicky Transient Facility (ZTF) is a 48-inch Schmidt telescope with a 47 sq. deg field of view at the Palomar Observatory in California. This large field of view ensures that the ZTF survey can scan the entire northern sky every night. The ZTF survey started on 2018 March 17. During the planned three years survey, ZTF is expected to acquire $\sim$ 450 observational epochs for 1.8 billion objects. Its main scientific goals are the physics of transient objects, stellar variability, and solar system science (Graham et al. 2019 [3]; Mahabal et al. 2019 [5]).

In this study we made an attempt to classify the first data release of ZTF survey (DR1) which contains data acquired between 2018 March and 2018 December, thus covering a timespan of around 290 days. The first data release includes more than 800 thousand light curves observed in both zr and zg passbands. We considered the machine learning technique as a perspective approach for the classification task. Our pre-processing procedure included several steps. First, we prepared the datasets with labels to use them later for training the model and testing the accuracy of the method. For this purpose the General Catalogue of Variable Stars was used. After cross-matching the catalog with

ZTF DR1 we found 19k common objects in zg-band, 14k in zr-band, and 13k in combination of passbands. Then, we chose the appropriate features to describe the light curves. As a starting point, we tried the magnitude amplitude range, the main peak period and power of Lomb–Scarge periodogram.

There are many types of variable stars differ by the underlying physical processes or their observational appearance. As the objects of interest we chose RR Lyrae, Cepheid and δ Scuti. We applied binary classification technique to Cepheid stars. "One vs all" classification technique revealed highly accurate results on validation data, concretely 0.90–0.95 with ROC-AUC metrics. The work done is a preparatory step towards the further thorough machine learning classification of the variable stars in ZTF data.

Acknowledgments

K. Malanchev and M. Pruzhinskaya are supported by RBFR grant 20-02-00779. The authors acknowledge the support from the Program of Development of M. V. Lomonosov Moscow State University (Leading Scientific School "Physics of stars, relativistic objects and galaxies").

References

1. M V Pruzhinskaya, K L Malanchev et al.,2019, Anomaly detection in the Open Supernova Catalog, Monthly Notices of the Royal Astronomical Society, Volume 489, Issue 3, November 2019, Pages 3591–3608, https://doi.org/10.1093/mnras/stz2362.
2. Richards et al., 2011, On Machine-Learned Classification of Variable Stars with Sparse and Noisy Time-Series Data. The Astrophysical Journal. 733. 10. 10.1088/0004-637X/733/1/10.
3. Matthew J. Graham et al.,2019, The Zwicky Transient Facility: Science Objectives, Published in: Publ.Astron.Soc.Pac. 131 (2019) 1001, 078001.
4. Samus N.N., Kazarovets E.V., Durlevich O.V., Kireeva N.N., Pastukhova E.N., General Catalogue of Variable Stars: Version GCVS 5.1, Astronomy Reports, 2017, vol. 61, No. 1, pp. 80–88
5. Ashish Mahabal et al., 2019, Machine Learning for the Zwicky Transient Facility. The Astronomical Society of the Pacific , Volume 131, Number 997.
6. Chen et al. 2020. The Zwicky Transient Facility Catalog of Periodic Variable Stars.
7. LombN.R. Least-squares frequency analysis of unequally spaced data. Astrophys Space Sci 39, 447–462 (1976). https://doi.org/10.1007/BF00648343
8. Scargle, J. D. 1982, Studies in astronomical time series analysis. II. Statistical aspects of spectral analysis of unevenly spaced data, Astrophysical Journal, Vol. 263, p. 835–853 (1982)
9. https://asas-sn.osu.edu/
10. https://www.h-atlas.org/
11. http://nesssi.cacr.caltech.edu/
12. https://gea.esac.esa.int/archive/
13. http://www.astropy.org

INFORMATION EXTRACTION FROM TEXT I

Exploring Book Themes in the Russian Age Rating System: a Topic Modeling Approach* (Extended Abstract)

Anna Glazkova[0000−0001−8409−6457]

University of Tyumen, Tyumen 625003, Russia a.v.glazkova@utmn.ru

Abstract. Age rating systems are to indicate target ages of potential content users based on information security and text semantics. Age ratings are usually given as a number, which tells us the youngest age the content is suitable for. A book or film with a 12+ rating has content which is suitable only for people aged 12 years and over, and a book or film with an 18+ rating is suitable for adults only. Currently, content assessment in terms of information security is carried out by experts. In this paper, we empirically compare book abstracts assigned to different age ratings using unsupervised topic modeling. We use an LDA model to discover topics from a collection of book abstracts. We then use statistical methods to study relations between the age rating categories assigned to books by experts and the topics obtained. We believe that our comparisons show interesting and useful findings for age rating automation. The full version of the article is placed in CEUR-WS volume of DAMDID/RCDL'2020 Proceedings.

Keywords: Topic modeling · Age restrictions· Age rating · Text classification · Statistical methods.

1 Introduction

The article deals with the comparing topics of texts assigned to different age rating categories according to the Russian Age Rating System (the RARS). The RARS contains 5 categories of content: (a) for children under the age of six (0+), (b) for children over the age of six (6+), (c) for children over the age of twelve (12+), (d) for children over the age of sixteen (16+), and (e) prohibited for children (18+).

Our findings can potential benefit many text classification applications, such as recommender systems and text filtering systems.

2 Methodology

We use a collection of abstracts for books in Russian. These abstracts was collected on the basis of public online libraries. Text preprocessing included both

* Supported by the grant of the President of the Russian Federation no. MK-637.2020.9.

standard steps, for example, converting to lower case, [14] and special actions for abstract preprocessing, such as removing personal names and words with low TF-IDF weights. This allowed us to exclude words typical of book abstracts ("book", "author", "reader", etc.) and to form themes according to the semantic proximity of abstracts, and not according to the belonging of books to one author or the coincidence of the characters names. Finally, we have combined common phrases (with a frequency of mutual occurrence of more than 5) into bigrams using the Gensim library [12].

To discover topics from the collection of abstracts, we built a topic model based on standard Latent Dirichlet Allocation (LDA) [15] with 100 topics. LDA approaches are widely used in topic modeling to analyze various subject areas, such as social media analysis [17,19,18], analysis of emails [5], news [7,16], fiction texts [10], and others.

Further, we calculated the topic distribution for each document in the collection. The topic distribution vector shows how a document extracted from the collection of abstracts corresponds to each topic. Then, we got the averaged topic distribution vector for each age rating category to obtain the average values of topic distribution for a group of documents. Applying the three sigma rule and the Dixon's Q-test [2] for our averaged vectors, we have highlighted the most typical topics for each age category and age-specific topics that are typical mainly for one age rating category.

More detailed statistics on the dataset and a description of the methodology are presented in the full article.

3 Empirical Analysis of Topics

According to the Russian law [3], books for children under 6 years old (0+) may contain episodic unnaturalistic images justified by the genre or descriptions of physical or psychological violence, provided that the victim is compassionate and happy ending. Our results showed that the prevailing topics in the 0+ category are fairy tales of the world, developing children's benefits, poems about the world around us and Christian literary works for children.

In books for children over the age of 6 (6+), non-naturalistic images or descriptions of human diseases, accidents, catastrophes or violent death without demonstrating their consequences are permissible. Books for children over 12 (12+) may contain scenes of violence or murder, descriptions of illnesses, disasters, but without details. Alcohol, tobacco and drug use may be present, but should be condemned. A schematic description of the hugs and kisses of men and women may be present. These two categories are described by similar topics in our topic model. These are short stories and tales for primary and secondary school age, fairy tales of the world, study guides and poems for children.

Books for children over 16 (16+) may contain scenes of illnesses, disasters without detailed descriptions. Violence, alcohol and drug use can be described, but should be condemned. Rough words may be present, with the exception of swear words. Scenes of sexual relations cannot be described with anatomical

details. In our example, this category is represented by military and human condition fiction, teaching aids, psychological and pedagogical literature.

A book should be marked with the 18+ label if the book contains a naturalistic description of illnesses, disasters, non-condemned drug and alcohol use, naturalistic scenes of sexual relations, non-traditional relationships, obscene language, scenes that encourage suicide. In our topic model, love stories, horoscopes, human condition fiction and possibly relationship psychology books were detected in this category.

Analyzing the results of topic modeling, we noticed that documents in category 0+ are largely mono-thematic. At the same time documents of other categories are usually mixtures of topics. Therefore, our topic model has many specific topics for texts from the 0+ category.

As it would be logical to assume, age-specific topics generally relate to children's books, as well as to specific literature from the 18+ category (in our case, literature on business and success).

A detailed analysis of the most common and age-specific topics is presented in the full version of the article.

4 Conclusion

In this paper, we empirically analyzed the topics of texts assigned to different age rating categories. We introduced the distribution of topics for age categories and the list of the most common topics for categories and age-specific topics. These list of topics were obtained using statistical methods. Our analysis confirmed the existing differences between the categories and demonstrated that topic models can be a good source of features for age rating identification. In our future work, we will try to develop a machine learning classifier for automatically determining the text age rating.

References

1. Blei, D. M., Ng, A. Y., Jordan, M. I. Latent dirichlet allocation. In: Journal of machine Learning research. Vol. 3(Jan). Pp. 993-1022 (2003).
2. Dixon, W. J.: Processing data for outliers. Biometrics **1**(9), 74–89 (1953). https://doi.org/10.2307/3001634
3. Federal Law of December 29, 2010 N 436-FZ (as amended on May 1, 2019) ¡¡On the Protection of Children from Information Harmful to Their Health and Development¿¿ (as amended and additional, entered into force on October 29, 2019) [Federal'nyj zakon ot 29.12.2010 N 436-FZ (red. ot 01.05.2019) ¡¡O zashchite detej ot informacii, prichinyayushchej vred ih zdorov'yu i razvitiyu¿¿ (s izm. i dop., vstup. v silu s 29.10.2019).], http://www.consultant.ru/document/cons_doc_LAW_108808/. Last accessed 7 Apr 2020.
4. Glazkova, A., Kruzhinov, V., Sokova, Z.: Dynamic Topic Models for Retrospective Event Detection: A Study on Soviet Opposition-Leaning Media. In: International Conference on Analysis of Images, Social Networks and Texts, pp. 145-154, Springer, Cham (2019). https://doi.org/10.1007/978-3-030-37334-4_13

5. Gong, H., You, F., Guan, X., Cao, Y., Lai, S.: Application of LDA Topic Model in E-Mail Subject Classification. In: 2018 International Conference on Transportation & Logistics, Information & Communication, Smart City. Atlantis Press (2018). https://doi.org/10.2991/tlicsc-18.2018.24
6. How are age-based gaming ratings set?, https://www.kaspersky.com/blog/gaming-age-ratings/11647/. Last accessed 7 Apr 2020.
7. Hu X.: News hotspots detection and tracking based on LDA topic model. In: 2016 International Conference on Progress in Informatics and Computing (PIC). IEEE, pp. 248-252 (2016). https://doi.org/10.1109/pic.2016.7949504
8. Ilyasova, R. S.: Dialectal lexis of P. P. Bazov's narrations ¡¡Malachite casket¿¿. Letters of the Chechen State University **3**(11), 103-107 (2018).
9. Manning, C.D.; Raghavan, P.; Schutze, H.: Scoring, term weighting, and the vector space model. Introduction to Information Retrieval. p. 100 (2008). https://doi.org/10.1017/CBO9780511809071.007.
10. Mitrofanova, O., Sedova, A.: Topic Modelling in Parallel and Comparable Fiction Texts (the case study of English and Russian prose). In: Proceedings of the International Conference IMS-2017, pp. 175-180 (2017). https://doi.org/10.1145/3143699.3143734
11. Raschka, S.: A questionable practice: Dixon's Q test for outlier identification, https://sebastianraschka.com/Articles/2014_dixon_test.html. Last accessed 13 Apr 2020. https://doi.org/10.13140/2.1.3000.0004
12. Rehurek, R., Sojka, P.: Gensimstatistical semantics in python, Retrieved from genism. org. (2011).
13. Natasha - high quality compact solution for extracting named entities from news articles in Russian, https://natasha.github.io/ner/. Last accessed 26 Jul 2020.
14. Loper, E., Bird, S.: NLTK: the natural language toolkit, arXiv preprint cs/0205028 (2002).
15. Vorontsov, K., Potapenko, A.: Tutorial on probabilistic topic modeling: Additive regularization for stochastic matrix factorization. In: International Conference on Analysis of Images, Social Networks and Texts, pp. 29-46, Springer, Cham (2014). https://doi.org/10.1007/978-3-319-12580-03.
16. Wang, H., Wang, J., Zhang, Y., Wang, M., Mao, C.: Optimization of Topic Recognition Model for News Texts Based on LDA. Journal of Digital Information Management **5**(17), 257 (2019). https://doi.org/10.6025/jdim/2019/17/5/257-269
17. Yang, S. and Zhang, H.: Text mining of Twitter data using a latent Dirichlet allocation topic model and sentiment analysis. In: Int. J. Comput. Inf. Eng, 12, pp.525-529 (2018).
18. Zhao, F., Zhu, Y., Jin, H., Yang, L. T.: A personalized hashtag recommendation approach using LDA-based topic model in microblog environment //Future Generation Computer Systems **65**, 196–206 (2016). https://doi.org/10.1016/j.future.2015.10.012
19. Zhao, W. X., Jiang, J., Weng, J., He, J., Lim, E. P., Yan, H., Li, X.: Comparing twitter and traditional media using topic models. In: European conference on information retrieval, pp. 338-349. Springer, Berlin, Heidelberg (2011). https://doi.org/10.1007/978-3-642-20161-53_4

Part of speech and gramset tagging algorithms for unknown words based on morphological dictionaries of the Veps and Karelian languages* (Extended Abstract)

Andrew Krizhanovsky[1,2][0000−0003−3717−2079], Natalia Krizhanovskaya[1][0000−0002−9948−1910], and Irina Novak[3][0000−0002−9436−9460]

[1] Institute of Applied Mathematical Research
of the Karelian Research Centre of the Russian Academy of Sciences
[2] Petrozavodsk State University
[3] Institute of Linguistics, Literature and History
of the Karelian Research Centre of the Russian Academy of Sciences,
Petrozavodsk, Russia, andrew.krizhanovsky@gmail.com
http://dictorpus.krc.karelia.ru

Abstract. This research devoted to the low-resource Veps and Karelian languages. Algorithms for assigning part of speech tags to words and grammatical properties to words are presented in the article. These algorithms use our morphological dictionaries, where the lemma, part of speech and a set of grammatical features (gramset) are known for each word form. The algorithms are based on the analogy hypothesis that words with the same suffixes are likely to have the same inflectional models, the same part of speech and gramset. The accuracy of these algorithms were evaluated and compared. 313 thousand Vepsian and 66 thousand Karelian words were used to verify the accuracy of these algorithms. The special functions were designed to assess the quality of results of the developed algorithms. 92.4% of Vepsian words and 86.8% of Karelian words were assigned a correct part of speech by the developed algorithm.

Keywords: Morphological analysis · Low-resource language · Part of speech tagging.

1 Introduction

Our work is devoted to low-resource languages: Veps and Karelian. These languages belong to the Finno-Ugric languages of the Uralic language family. Most Uralic languages still lack full-fledged morphological analyzers and large corpora [5].

* The study was supported by the Russian Foundation for Basic Research, grant 18-012-00117.

Our Open corpus of Veps and Karelian languages (VepKar) contains morphological dictionaries of the Veps language and the three supradialects of the Karelian language: the Karelian Proper, Livvi-Karelian and Ludic Karelian. The developed software (corpus manager)[4] and the database, including dictionaries and texts, have open licenses.

Algorithms for assigning part of speech tags to words and grammatical properties to words, without taking into account a context, using manually built dictionaries, are presented in the article (see Section 2). The evaluatation of accuracy of the algorithms is presented in the Section 3.

Let us describe several works devoted to the development of morphological analyzers for the Veps and Karelian languages.

- The Giellatekno language research group is mainly engaged in low-resource languages, the project covers about 50 languages [4]. Our project has something in common with the work of Giellatekno in that (1) we work with low-resource languages, (2) we develop software and data with open licenses.

  A key role in the Giellatekno infrastructure is given to formal approaches (grammar-based approach) in language technologies. They work with morphology rich languages. Finite-state transducers (FST) are used to analyse and generate the word forms [4].
- There is a texts and words processing library for the Uralic languages called UralicNLP [2]. This Python library provides interface to such Giellatekno tools as FST for processing morphology and constraint grammar for syntax. The UralicNLP library lemmatizes words in 30 Finno-Ugric languages and dialects including the Livvi dialect of the Karelian language.

2 Part of speech and gramset search by analogy algorithms

The proposed algorithms operate on data from a morphological dictionary. The algorithms are based on the analogy hypothesis that words with the *same suffixes* are likely to have the same inflectional models and the same sets of grammatical information (part of speech, number, case, tense, etc.). The *suffix* here is a final segment of a string of characters.

Let the hypothesis be true, in that case, if the suffixes of new words coincide with the suffixes of dictionary words, then part of the speech and other grammatical features of the new words will coincide with the dictionary words. It should be noted that the length of the suffixes is unpredictable and can be various for different pairs of words [1, p. 53].

[4] See https://github.com/componavt/dictorpus

2.1 The POSGuess algorithm for part of speech tagging with a suffix

Given the set of words W, for each word in this set the part of speech is known. The POSGuess algorithm finds a part of speech pos_u for a given word u using this set.

In POSGuess algorithm we look for in the set W the words which have the same suffix u_z as the unknown word u. Firstly, we are searching for the longest substring of u, that starts at index z. The first substring $u_{z=2}$ will start at the second character, since $u_{z=1} = u$ is the whole string.

Then we increment the value z decreasing the length of the substring u_z in the loop, while the substring u_z has non-zero length, $z \leq len(u)$. If there are words in W with the same suffix, then we count the number of similar words for each part of the speech and stop the search.

2.2 The GramGuess algorithm for gramset tagging with a suffix

The GramGuess algorithm is exactly the same as the POSGuess algorithm, except that it is needed to search a subset of gramsets instead of parts of speech. That is in the set W the gramset is known for each word. The gramset is a set of morphological tags (number, case, tense, etc.).

3 Experiments

3.1 Part of speech search by a suffix (POSGuess algorithm)

For the evaluation of the quality of results of the searching algorithm POSGuess the following function eval(pos^u) was proposed:

$$\text{eval}\left(\begin{matrix}\text{pos}^u,\\ \text{Counter}\left[\text{pos}^k\right] \to c^k,\\ \forall k = \overline{1,m}\end{matrix}\right) = \begin{cases} \text{The array Counter[] do not contain the correct pos}^u. \\ 0,\ \text{pos}^u \neq \text{pos}^k, \forall k = \overline{1,m}, \\ \text{First several POS in the array can have the same} \\ \text{maximum frequency } c^1\text{, one of this POS is pos}^u. \\ 1,\ \text{pos}^u \in \{\left[\text{pos}^1, \ldots, \text{pos}^j\right] : c^1 = c^2 = \ldots = c^j, \\ \qquad\qquad j \leq m\}, \\ \dfrac{c^k}{\sum_{k=1}^{m} c^k},\ \exists k : \text{pos}^k = \text{pos}^u,\ c^k < c^1 \end{cases} \tag{1}$$

This function (1) evaluates the result of the POSGuess algorithm against the correct part of speech pos^u. The POSGuess algorithm counts the number of words similar to the word u separately for each part of speech and stores the result in the Counter array.

71,091 “word – part of speech” pairs for the Karelian Proper supradialect and 399,260 “word – part of speech” pairs for the Veps language have been used in the experiments to evaluate algorithms.

Figure 1 shows the proportion of Veps and Karelian words with correct and wrong part of speech assignment by the POSGuess algorithm. Values along the X axis are the values of the function eval(pos^u), see the formula (1).

92.38% of Vepsian words and 86.77% of Karelian words ($x = 1$ in Fig. 1) were assigned the correct part of speech.

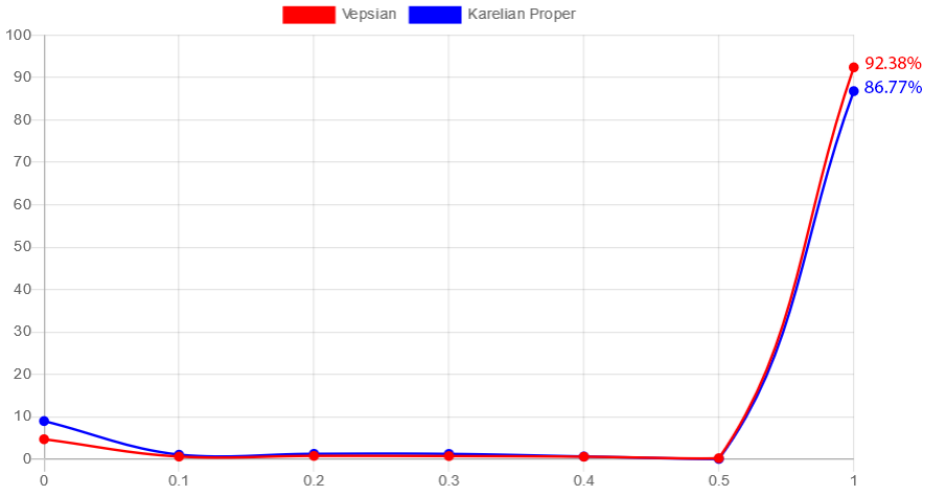


Fig. 1. The proportion of Vepsian (red curve) and Karelian (blue curve) words with correct ($x = 1$) and wrong ($x = 0$) part of speech assignment by the POSGuess algorithm with the formula (1).

3.2 Gramset search by a suffix (GramGuess algorithm)

A list of gramsets was searched for each word. The list was ordered by the number of similar words having the same gramset.

90.7% of 73,395 Karelian words and 95.3% of 452,790 Vepsian words were assigned a correct gramset by GramGuess algorithm.

References

1. G. G. Belonogov, Yu. P. Kalinin, A. A. Khoroshilov: Computer Linguistics and Advanced Information Technologies: Theory and Practice of Building Systems for Automatic Processing of Text Information (In Russian). Russian World, Moscow (2004)

2. Hämäläinen, M.: UralicNLP: An NLP Library for Uralic Languages. Journal of open source software, **4**(37), 1345 (2019). https://doi.org/10.21105/joss.01345
3. Klyachko, E. L., Sorokin, A. A., Krizhanovskaya, N. B., Krizhanovsky, A. A., Ryazanskaya, G. M.: LowResourceEval-2019: a shared task on morphological analysis for low-resource languages. In: Conference "Dialog", 45–62. Moscow, Russia (2019). arXiv:2001.11285.
4. Moshagen, S., Rueter, J., Pirinen, T., Trosterud, T. and Tyers, F.M.: Open-source infrastructures for collaborative work on under-resourced languages. In: Collaboration and Computing for Under-Resourced Languages in the Linked Open Data Era, 71–77. Reykjavík, Iceland (2014).
5. Pirinen, T. A., Trosterud, T., Tyers, F. M., Vincze, V., Simon, E., Rueter, J.: Foreword to the Special Issue on Uralic Languages. Northern European Journal of Language Technology **4**(1), 1–9 (2016). https://doi.org/10.3384/nejlt.2000-1533.1641

Extrinsic evaluation of cross-lingual embeddings on the patent classification task (Extended Abstract)

Anastasiia Ryzhova[1] and Ilya Sochenkov[2]

[1] Skolkovo Institute of Science and Technology, Moscow, Russia
Anastasiia.Ryzhova@skoltech.ru
[2] Federal Research Center "Computer Science and Control" of the Russian Academy of Sciences, Moscow, Russia sochenkov@isa.ru

Keywords: patent classification · cross-lingual embeddings · zero-shot cross-lingual classification · MUSE embeddings · LASER embeddings · Multilingual Bert · XLM-RoBERTa.

The patent is a legal document that helps its holder to protect the new invention from selling it by other people. Each patent document has a specific structure, including an abstract, detailed description, summary of the invention, drawings, and claims. There exists a universal classification system called the International Patent Classification (IPC). It allows classifying patent documents in different languages. This system has a hierarchical structure; the lowest levels provide more in-depth specificity.

It is difficult to assign the IPC code manually, and many researchers attempted to develop automatic tools for solving this problem ([6], [7], [8], [2]).

In our work, we performed the automatic classification of patent documents at the subclass level (the first four digits of the code). However, the improvement of existing patent classification methods was not the primary goal. The main aim was to compare the quality of cross-language embeddings (extrinsic evaluation) on zero-shot cross-lingual classification task, which means that training and test sets include documents in different languages. In our experiments we used 80,948 patent documents in the English language, which we divided into training and validation sets in the ratio of 85% / 15%. These sets include patent documents of 225 different subclasses. As the test set, we considered 10,150 texts in Russian of the same subclasses. We used only the abstract section of the patent document, which contains concise information about the invention in minimal number of words ($\approx$ 100 words).

The primary purpose of cross-lingual representations is to compare the meanings of words in different languages. Also, the universal representations help to analyze and solve the NLP tasks in low-resource languages by transferring the knowledge from rich-resource ones. We considered four cross-lingual word and sentence representations: Multilingual Unsupervised and Supervised Embeddings (MUSE, [4,9]), Language-Agnostic SEntence Representations (LASER, [1]), Multilingual Bert model (MBert, [5]), and XLM-RoBERTa model (XLM-R,

[3]). In the first two cases we used the embeddings as the inputs for various machine and deep learning models. The Multilingual Bert and XLM-RoBERTa models were fine-tuned on our patent classification task, for this purpose we added the linear layer on the top of the last pooling layer of models. In MBert case, for better performance, we froze the embedding layer to make the model less tuned to the English language.

We do not have a high imbalance in our data, and in this task it is not so significant to have high accuracy on small classes, so we decided to use the f1-micro and f1-weighted scores as evaluation metrics. Also, we compared how the performance of the embeddings depends on the language. For this reason, we translated the Russian patent documents in English with the Yandex translator. This translated test dataset we denote as "Test translated". Table 1 presents the main results of our experiments:

Table 1. Results of main experiments, f1-micro/f1-weighted scores

Model	Train	Validation	Test	Test translated
MUSE + SVM	48.63/47.53	44.92/43.52	12.81/11.97	38.70/37.80
MUSE + Logistic regression	44.40/42.80	42.10/40.24	19.87/17.27	37.54/34.98
MUSE + Feed-forward neural network	43.92/42.35	41.92/40.17	17.48/15.26	38.56/36.28
MUSE + BiLSTM with attention	55.33/54.42	51.54/50.63	**25.36/23.41**	47.75/46.14
MUSE + CNN	60.50/59.89	51.67/50.98	22.62/20.97	44.95/43.88
LASER + SVM	56.33/55.70	43.57/42.05	28.78/28.34	38.74/37.05
LASER + Logistic regression	54.76/53.85	43.87/42.16	31.72/29.56	39.69/37.84
LASER + Feed-forward neural network	56.46/55.49	45.22/44.01	**32.82/31.14**	40.38/38.69
MBert	71.96/70.50	**61.62/60.15**	25.05/21.81	**59.52/57.68**
XLM-Roberta	70.43/68.84	61.44/59.84	**41.06/38.69**	**59.76/57.82**

We can conclude that the XLM-R model embeddings outperform other approaches and show a better ability to transfer knowledge between languages. This model was pre-trained on a large amount of data (more than two terabytes of filtered CommonCrawl data), and better catches the semantic similarity between words of domain-specific patent vocabulary on different languages.

In future it will be interesting to hold experiments using test data on other languages. Also we want to combine different embeddings in one ensemble model and see the performance.

Acknowledgments

The reported study was funded by RFBR according to the research projects No. 17-29-07088 and No. 18-37-20017.

References

1. Artetxe, M., Schwenk, H.: Massively multilingual sentence embeddings for zero-shot cross-lingual transfer and beyond (12 2018)
2. Chen, Y.L., Chang, Y.C.: A three-phase method for patent classification. Information Processing & Management **48**, 1017–1030 (11 2012). https://doi.org/10.1016/j.ipm.2011.11.001
3. Conneau, A., Khandelwal, K., Goyal, N., Chaudhary, V., Wenzek, G., Guzman, F., Grave, E., Ott, M., Zettlemoyer, L., Stoyanov, V.: Unsupervised cross-lingual representation learning at scale (11 2019)
4. Conneau, A., Lample, G., Ranzato, M., Denoyer, L., Jégou, H.: Word translation without parallel data. arXiv preprint arXiv:1710.04087 (2017)
5. Devlin, J., Chang, M.W., Lee, K., Toutanova, K.: Bert: Pre-training of deep bidirectional transformers for language understanding (10 2018)
6. Fall, C., Trcsvri, A., Benzineb, K., Karetka, G.: Automated categorization in the international patent classification. SIGIR Forum **37**, 10–25 (04 2003). https://doi.org/10.1145/945546.945547
7. Fall, C., Trcsvri, A., Fivet, P., Karetka, G.: Automated categorization of german-language patent documents. Expert Systems with Applications **26**, 269–277 (02 2004). https://doi.org/10.1016/S0957-4174(03)00141-6
8. Kim, J.h., Choi, K.S.: Patent document categorization based on semantic structural information. Information Processing & Management **43** (09 2007). https://doi.org/10.1016/j.ipm.2007.02.002
9. Lample, G., Conneau, A., Denoyer, L., Ranzato, M.: Unsupervised machine translation using monolingual corpora only. arXiv preprint arXiv:1711.00043 (2017)

Automated Generation of a Book of Abstracts for Conferences that use Indico Platform (Extended Abstract)

Anna Ilina[1][0000−0002−2498−2091] and Igor Pelevanyuk[1,2][0000−0002−4353−493X]

[1] Joint Institute for Nuclear Research, Dubna, Russia
[2] Plekhanov Russian University of Economics, Moscow, Russia

Keywords: Document generation · Book of abstracts · Automated system · DOCX

Automatic document generation is an important topic especially for systems that process a big amount of data to generate standardized reports or documents. Manual document creation may be tedious and error-prone. And if the data amount is large and document format is not trivial the creation of a document may become too expensive in terms of time. Without some additional efforts there is no way to prove that the document is free of typos or mistakes.

The task of document generation is important for Joint Institute for Nuclear Research(JINR) - International Intergovernmental Organization which unites scientists from areas of nuclear physics, high-energy physics, neutron physics, information technologies, and radiation biology. JINR organizes and hosts more than 40 international conferences and meetings annually. Organization of a conference requires a substantial amount of work related to conference information publication, registration of participants, abstracts collection, time table creation etc. For this purpose, the Indico system is widely used.

Indico is a software started as a European project in 2002. In 2004, CERN adopted it as its own event-management solution and has financed its development since. Today Indico is an Open Source Software available under MIT license. As a tool for conference organization, Indico provides a web page with information about the event, registration form, abstract submission form, survey form, time schedule of the event, links to web pages with other information related to the event, and some other features. With its rich functionality, in the world of high energy physics, Indico became a de facto standard tool for event organization[1].

The book of abstracts is an important aspect of a scientific conference. It should be accessible before the talks start. The book of abstracts may appear in two forms: a printed copy or a digital copy. Standard abstract consists of several blocks of information: title, list of authors, list of affiliations, corresponding author email, and the text of the abstract itself.

Each author has one or several affiliations. Affiliations marked by numbers. The email address relates to the corresponding author. Sometimes there are several corresponding authors. Email addresses marked by letters.

Our primary goal was to simplify a process of a book of abstracts creation for the International Symposium on Nuclear Electronics and Computing(NEC) organized by JINR in 2019. Usually, a book of abstract for this conference consists of around 150 abstracts. The creation of this book was performed manually and required a lot of effort. The work was monotonous, boring, and related to many copy/paste cycles. That led to typos and errors during document creation. The work could be divided between several people which lead to spending of around 30 man-hours in total just for one book of abstracts. Manual indexing of affiliations could lead to mistakes and required from the editor an additional check.

Important issue during the creation of a book of abstracts is the fact that if we just put all of the abstracts from abstract submission forms in one document the inconsistencies and errors will become visible. Among all the different problems the following are more common:

1. Authors with the same affiliation writes them differently. For example: JINR, LIT JINR, Joint Institute for nuclear research, etc.
2. Abstract titles in the book of abstracts may be either fully capitalized or just starting with the capital. The editor of the book of abstract should check every title manually and bring it to the right format.
3. The languages of different pieces of information about abstracts may be written in a different language. Generally, it requires organizers to communicate with the author to get the right spelling of the name in the language of the abstract text.
4. Abstract text is usually confined by 250 words. The editor sometimes should manually check the number of words and send requests to authors to shorten texts.

The described requirements and issues demonstrate that the problem of the generation of a document with some specific format is not the only issue with document generation. The analysis and automatic corrections of the text are also possible and required. It is possible to correct the affiliation name or capitalize on the title of an abstract. But, mixing of languages inside an abstract usually requires organizers to communicate with authors. To perform automatic analysis of texts for book abstracts it is required to have a convenient source of data. Fortunately, Indico provides the possibility to download the XML document with the representation of all authors and abstracts. In the newer versions of the Indico, the XML format has been replaced with JSON. In the scope of this article, we will refer only to XML since it was the only option available for use at the moment.

Using the data to generate the final document automatically would give great flexibility in terms of content and the form of the final document. The generation of the PDF file directly from the data is not as simple as a generation of PDF from HTML, DOCX, or TEX formats. We will overview just two common ways to generate PDF: from TEX file using the LaTeX document preparation system, and from DOCX file using Microsoft Word word processor.

LaTeX is the standard for the publication of scientific documents. It provides high-quality document printing, so the document looks like a book.

The biggest advantage of this method is the possibility to generate a TEX file without third-party libraries. Once the template TEX file is available it is rather easy to fill it with text form source XML file. And LaTeX system itself is free software.

However, the preparation of a template in the LaTeX system usually requires some special skills. Another issue with LaTeX was the fact that not all organizing committees had a LaTeX template available. Some committees used a complex DOCX template with macros to apply correct formatting to different parts of a text. Also, not all LaTeX editors support the WYSIWIG mode. Some editors that support this mode may be non-free.

Another approach to generate PDF is to use a prepared DOCX template. Templates may be different. It is possible to make it simple and just define several types of formatting for different fields, like Abstract title, Authors, Abstract text. The resulting document may be generated from a template using special third-party libraries. The generated DOCX document may be additionally edited in Microsoft Word. The WYSIWIG is originally supported by Microsoft Word and may simplify the manual editing process.

The disadvantage of this approach is the need to use third-party libraries. There is no guarantee that they will always generate the correct document and the newer version of Microsoft Word may render generated DOCX files differently. Microsoft Word itself is non-free software, however, currently, it is a standard software for document creation in many organizations, including JINR.

There were three reasons for us to use DOCX templates for the generation of the book of abstracts. First, the DOCX template has already been created and used for several previous NEC conferences. Second, libraries for Python language for work with DOCX templates and documents have been found, and their functionality was proved during initial tests. Third, in our case, members of the organizing committees responsible for the books of abstracts preferred Microsoft Office.

The Python3 language has been chosen as a primary language for the developed system. It is possible to use programs developed in Python under Linux and Windows operating systems. Moreover, Python provides the possibility to make the developed program available as a web-application. The Python is quite popular in science organizations and the developed system may be easily used and changed by other users.

The generation of a book of abstracts is done in several steps:

1. Export of the XML file with all abstracts data.
2. Parsing of XML file in Python and creating object with all data from XML file.
3. Automatic correction of standard inconsistencies.
4. Displaying the notifications about issues that cannot be fixed automatically.
5. Generation of the preface part with general conference information.
6. Generation of abstracts part which contain only abstracts.

7. Creation of the final document by concatenating preface and abstract parts.

To allow the execution of these steps several additional files are required: templates, information about the conference to be used for preface generation, CSV file with validated affiliation names.

The work involved designing and creating an automated abstract book generator that would be identical to the one created by hand. Besides, we had a task to check the source XML file for possible errors listed in this article. As a result, a software product was developed. The schema of its work is on the Fig. 1

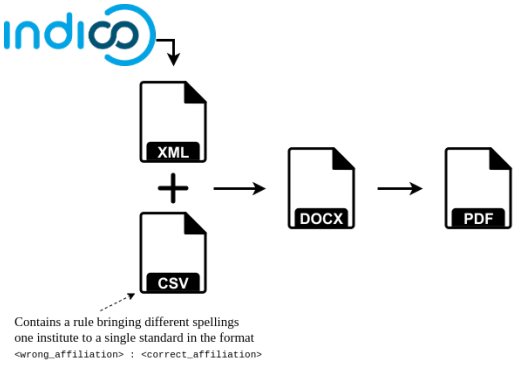


Fig. 1. Steps for converting input information to an output file.

The developed program creates a final DOCX file with the formatting used in the template. Besides, the program performs checks for possible errors described in this article. Using the information obtained during checks, an editor of the book of abstracts can decide how to correct mistakes: make changes in Indico, ask for a new feature for the book of abstracts generation tool, or change the DOCX after generation manually. The applied approach greatly reduced the amount of effort required to generate a document of a book of abstracts. The pursuit to simplify manual indexing and copying led to introducing automatic checks and corrections. And developed system allowed to generate a document in a format that is known to the end-user. The developed product is a command-line interface (CLI) application written in Python v3.5 for Linux and Windows. This requires some specific libraries to be installed on the client machine before the use of the system. The possible next step is providing a developed system as a service through a Web Interface. All source code is available under GNU General Public License v3.0 at [2]. The software product, methods, and approaches described in this article were used during the preparation of the book of abstracts for the NEC2019 Symposium. The generated book is available at [3].

References

1. Indico Project , https://docs.getindico.io/en/latest/. Last accessed 25 July 2020
2. Project GitHub repository https://github.com/trnkv/IndicoAbstract. Last accessed 25 July 2020
3. Generated book of abstracts for NEC2019 conference https://indico.jinr.ru/event/738/attachments/4884/6443/NEC_2019_BoA.pdf accessed 25 July 2020

Text Attribution in Case of Sampling Imbalance by the Method of Constructing an Ensemble of Classifiers Based on Decision Trees (Extended Abstract) *

Alexander Rogov[1], Roman Abramov[1], Alexander Lebedev[1][0000−0001−9939−9389], Kirill Kulakov[1][0000−0002−0305−419X], and Nikolai Moskin[1][0000−0001−5556−5349]

Petrozavodsk State University, Petrozavodsk, Russia
rogov@petrsu.ru, monset008@gmail.com, perevodchik88@yandex.ru, kulakov@cs.karelia.ru, moskin@petrsu.ru
https://petrsu.ru/

Abstract. When solving the attribution problem, the question of determining the author's style of a writer who created a smaller number of texts (both quantitatively and in terms of the total number of words) in comparison with other analyzed authors arises. In this paper we consider possible solutions to this problem by the example of determining the style of Apollon Grigoriev. As a method for constructing an ensemble of classifiers we use *Bagging* (*Bootstrap aggregating*). The SMALT information system ("Statistical methods for analyzing literary texts") was used to determine the frequency characteristics of the texts and Python 3.6 was used to build decision trees. As a result of calculations we can assume that the relative frequency of the "particle-adjective" bigram more than 6.5 is a distinctive feature of the journalistic style of Apollon Grigoriev. There also was a study of the article "Poems by A. S. Khomyakov", which confirms the previously conclusion that there is no reason to consider it as belonging to Apollon Grigoriev.

Keywords: Text attribution · F. M. Dostoevsky · Apollon Grigoriev · Poems by A. S. Khomyakov · sampling imbalance · decision tree · software complex "SMALT".

1 Introduction

Authorship identification of anonymous texts (attribution of texts) is one of most urgent problem for the philological community [3]. One of the issues, which is far from its final decision, is the affiliation of anonymous articles published in the magazines "Time" and "Epoch" (1861-1865) [2]. The solution to this problem is additionally hampered by the uneven amount of available textual material: there are many articles owned by F. M. Dostoevsky, while the remaining authors

* Supported by the Russian Foundation for Basic Research, project no. 18-012-90026.

published in these journals (for example, A. Grigoriev, N. N. Strakhov, Ya. P. Polonsky, etc.), don't have so many texts that are uniquely attributed to them.

Different mathematical methods are used to establish authorship of works. Among them decision trees are distinguished by the fact that they are easy to understand and interpret and also do not require special preliminary data processing. When solving the problem of classification into two classes, the problem of sampling imbalance often arises, i.e. when the number of objects of one class significantly exceeds the number of objects of another class. In this case the first class is called the majority class and the second class is called the minority class. In such samplings classifiers are configured for objects of the majority class, i.e. high accuracy of the classifier can be obtained without selecting objects of the minority class. When solving the attribution problem, the question of determining the author's style of a writer who created a smaller number of texts (both quantitatively and in terms of the total number of words) in comparison with other analyzed authors arises.

2 Construction and analyzing decision trees

As a method for constructing an ensemble of classifiers we use *Bagging* (*Bootstrap aggregating*) [1]. The authors believe that it meets the meaning of the task better than *Boosting*. Based on previous research, the fragment size was chosen to be 1000 words and the step size for choosing the beginning of the next fragment to be 100 words. The SMALT information system was used to determine the frequency characteristics [3]. Specialists in philology carried out grammatical markup of texts, which took into account 14 parts of speech. A set of data for training was compiled (118 fragments – Apollon Grigoriev, 899 – the rest). In this case fragments of the texts of Apollon Grigoriev are objects of the minority class and all the others are from the majority class. The text size is quite small (from 2000 to 7000 words).

Python 3.6 was used to build decision trees (libraries: *scikit-learn* – for tree implementation, *pandas* – for data reading). The original data set was divided into 7 parts. All fragments of Apollon Grigoriev were taken as a class with a label "1", the same number of fragments of other authors were taken randomly as a class with a label "0". Repetitions of fragments of other authors were not allowed. A decision tree was trained on each part of data. The training continued until accuracy reached 100% (tree depth). All trees formed an ensemble. The decision was accepted by a majority vote. *Accuracy* was calculated on the entire data as $(TP+TN)/(TP+TN+FP+FN)$, where TP is true-positive, TN is true-negative, FP is false-positive and FN is false-negative predicted class. As a result of experiments depth 1 corresponds to the classifier accuracy of 0,8628 (respectively 2 – 0,9592, 3 – 0,9841, 4 – 0,9891, 5 – 0,992, 6 – 0,9901).

In total 7 decision trees were built. Note that on the third level there are two leaves that contain a small number of fragments (summary from 12 to 27, on average less than 8%). You should take into account the possible inaccuracy of the source data. The texts of Apollon Grigoriev could be edited by F. M. Dos-

toevsky. In addition there is a slight volatility in the parameters of the author's style depending on external factors (such as mood, health status, etc). Therefore, when solving the problem of text attribution, you should limit yourself to the first level or at most the first two levels of decision trees. The accuracy of the ensemble at the second level already falls into the generally accepted 5% significance level. Analyzing the decision trees contained in the ensemble, it can be noted that in 4 of them the first attribute was the "particle-adjective" bigram less than or equal to 6.5. In two cases the same attribute is found, but with a different threshold (less than or equal to 7.5). Only one tree had a different first attribute ("adjective-particle") less than or equal to 2.5.

The influence of the universally accepted methods for processing unbalanced data "UpSampling", "UnderSampling", "SMOTE" on the accuracy of classification of works by Apollon Grigoriev was analyzed. The available data set was divided into test (42 - Apollon Grigoriev, 310 - Other) and training samples. The training sample was subjected to the techniques listed above to confront class imbalance. Then the accuracy ("Accuracy", "roc-auc" curve) was calculated on a test sample, which was the same for all three techniques. This analysis showed approximately the same accuracy of all three methods. UpSampling looks worse. The advantage of UnderSampling is that it is easier to explain. Therefore, the authors decided to focus on it.

One of the controversial and still unresolved issues is the article "Poems by A. S. Khomyakov". This work has long been attributed to Apollon Grigoriev. However, recently it has been considered the copyright text of F. M. Dostoevsky [4]. It was interesting to check where our classifier will take it. The text will be attributed to the author that most of the text fragments belong to. If we take the classification on the first node, then 6 of the 7 decision trees classify it as "Other", i.e. as not the text of Apollon Grigoriev. Only on one tree, there was an equality (5 fragments "for belonging" and 5 "against"). During the split on the second level 3 "for belonging", 3 "against" and in one rejection of the classification. Our study confirms the earlier conclusion [4] that there is no reason to consider the article "Poems by A. S. Khomyakov" as belonging to Apollon Grigoriev.

3 Information system SMALT

The SMALT information system developed at Petrozavodsk State University is designed for the collective work of various specialists with texts [3]. The information system can be divided into three sections: import of new texts, verification of texts by philologists and the use of various analysis methods both on a single text and for a group of texts. As part of the text import process, the text is divided into sections, paragraphs, sentences and words, as well as matching each word with its morphological analysis. If the task of text separation is typical, then the task of comparing the morphological analysis is rather complicated. The problem is both in the wide variety of spelling of the word (using pre-revolutionary graphics, a more flexible dictionary allowing different spelling of the word), and

in the need to take into account the context of the use of the word. At different times, algorithms for finding the first possible variant, a frequently used variant and an algorithm based on n-grams were used to select the semantic analysis of the word. The latter has a great prospect due to the small number of subsequent corrections.

As part of the text verification process, philologists perform correction of text analysis (for example, combining or separating words), correction of morphological analysis of a word, or creation of a new analysis. Using the web interface allows several specialists to work on the text at the same time. During the analysis process, SMALT provides researchers with access to the accumulated database in various sections. For example, one of the popular statistical characteristics is Kjetsaa metrics [2]. SMALT calculates the characteristics of both a single work and a group of texts. Another objective of the analysis is to identify the causes of the results. For example, to identify the reasons for the separation of text fragments between different nodes of the decision tree. The SMALT information system allows you to access the source data of the required fragment for subsequent linguistic analysis.

4 Conclusion

When solving the problem of determining the author's style of Apollon Grigoriev, the problem of sampling imbalance often arises. Analyzing decision trees, we can assume that the relative frequency of the "particle-adjective" bigram more than 6.5 is a distinctive feature of the journalistic style of Apollon Grigoriev. The obtained knowledge was used to study the authorship of the article "Poems by A. S. Khomyakov", a discussion about whose authorship in the literary criticism continues over the past twenty years. If we take the classification on the first node, then 6 of the 7 decision trees classify it as "Other", i.e. as not the text of Apollon Grigoriev.

5 Acknowledgements

This work was supported by the Russian Foundation for Basic Research, project no. 18-012-90026.

References

1. Bühlmann, P.: Bagging, Boosting and Ensemble Methods. In: Gentle J., Härdle W., Mori Y. (eds) Handbook of Computational Statistics. Springer Handbooks of Computational Statistics. Springer, Berlin, Heidelberg (2012). https://doi.org/10.1007/978-3-642-21551-3_33
2. Kjetsaa, G.: Attributed to Dostoevsky: The Problem of attributing to Dostoevsky anonymous articles in Time and Epoch. Oslo: Solum Forlag A. S. (1986)

3. Rogov, A., Kulakov, K., Moskin, N.: Software support in solving the problem of text attribution. Software engineering **10**(5), 234–240 (2019) https://doi.org/10.17587/prin.10.234-240
4. Zakharov, V.: Question about Khomyakov. In: Zakharov, V. The name of the author is Dostoevsky. Essay on creativity. Moscow, Indrik, 231–247 (2013)

INFORMATION EXTRACTION FROM TEXT II

An Approach to Extracting Ontology Concepts from Requirements (Extended Abstract)

Marina Murtazina[1][0000-0001-6243-9308] and Tatiana Avdeenko[1][0000-0002-8614-5934]

[1] Novosibirsk State Technical University, 630073 Novosibirsk, Russia
murtazina@corp.nstu.ru

Abstract. The paper proposes an approach to extracting ontology concepts from the results of automatic processing the software requirements written in natural Russian-language. Relevance of developing an approach to automating the process of extraction of the ontology concepts from the requirements written in natural language is due to the necessity to maintain the requirements specification in a consistent state under conditions of business environment variability. First we discuss the advantages of the ontology-based approach to the requirements engineering as a stage of the software engineering process. The main attention is focused on the possibilities of presenting knowledge about the software application domain and requirements specification in the form of ontologies. We analyzed possibilities of automatic Russian text processing tools for extracting ontology concepts and consider such tools as ETAP-4, MaltParser and UDPipe. The choice of UDPipe as a tool for automatic processing the requirements texts is explained by the best results that it showed when analyzing texts such as software requirements. Then we describe the main classes, object properties and data properties of the proposed requirements ontology and the application domain ontology. An additional advantage of the proposed approach, increasing its practical utility, is building the system of production rules by which the analysis of the results of the automatic processing of the requirements texts obtained using the UDPipe tool.

Keywords: Requirements engineering · Ontology · Automatic text processing.

1 Analysis of automatic Russian text processing tools possibilities for extracting ontology concepts

During the research, the ETAP-4, MaltParser and UDPipe tools were considered for marking up the requirements presented in natural Russian-language. We first compared ETAP-4 and MaltParser. Parser results are almost the same. Therefore, the convenience of integration with other software systems was a decisive factor for choosing between these two tools for research purposes. Since MaltParser is a java application, its use becomes preferable because the execution of the application does not depend on the operating system, whereas ETAP-4 is Windows-based software. Since Malt-

Parser does not perform tokenization, morphological tagging and lemmatization, it is necessary to convert text data into one of the input formats before using it. As a part of this research, a language-independent TreeTagger tool (version 3.2.1) was used for morphological tagging and lemmatization of the text. For automatic processing of simple text files (files containing only text), a set of scripts in Python (version 3.7.0) and a batch file for execution of scripts under the Windows operating system were developed (a similar shell script can be made for Linux family of OSs).

The next tool, that MaltParser is compared to be, is trainable pipeline UDPipe. This tool performs tokenization, tagging, lemmatization and dependency parsing of CoNLL-U. As a part of the study of the possibilities of using the MaltParser and UDPipe tools for the purposes of this work, Russian language models were trained on the UD Russian-SynTagRus case (version 2.0-170801). After training both tools on the same data, a qualitative comparison of the results was carried out. It is worth noting that in MaltParser results for complex sentences sometimes incorrect results. MaltParser does not always correctly assign the tag ”root”. In addition, MaltParser requires pre-processing of the text, unlike UDPipe. In connection with the above and the ease of use of the UDPipe tool itself, the latter was chosen. For further work, we used the pre-trained Russian language model russian-syntagrus-ud-2.4-190531.udpipe, corresponding to the requirements of the Universal Dependencies (UD) framework. Using the UDPipe tool that supports the Russian language model that meets the requirements of the Universal Dependencies project, it is possible to develop a set of concept extraction rules that will be closer to existing English-language analogs of rules using syntactic dependencies.

2 An approach to extracting ontology concepts from the results of automatic processing of textual requirements

In this section, we will consider some classes of the application domain ontology that can be updated based on the proposed approach to extracting concepts, as well as some classes of requirements ontologies that are proposed to be filled using the developed approach. The domain area terminology is proposed to be presented taking into account some data properties, object properties and classes characteristic of lexical ontologies. The words that make up the concepts include nouns, verbs, auxiliary verbs, and definitions compatible with nouns. The ontological graph of classes ”DomainConcept” and ”Word” is shown in Fig. 1. In the class ”DomainConcept”, there are two main subclasses that are needed to describe of requirements: Entity and RelationAction. The requirements ontology is intended to analyze a set of requirements presented in the form of an ontology concepts. The process of extracting ontology concepts from the requirements for subsequent analysis is extreme-ly time consuming and requires automation. Previously unknown concepts can be ex-tracted into the requirements ontology from text requirements, then external WordNet-like resources can be used to identify relations between the con-cepts. Classes, object and data properties for requirements ontology are shown in Fig. 2.

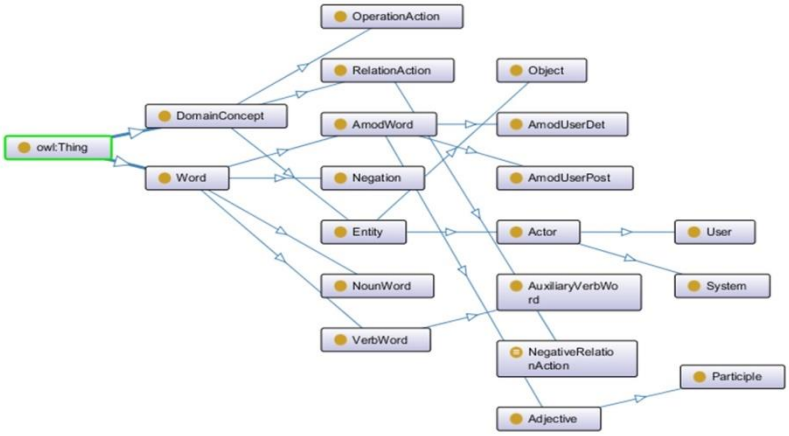


Fig. 1. Ontological graph of classes ”DomainConcept” and ”Word”.

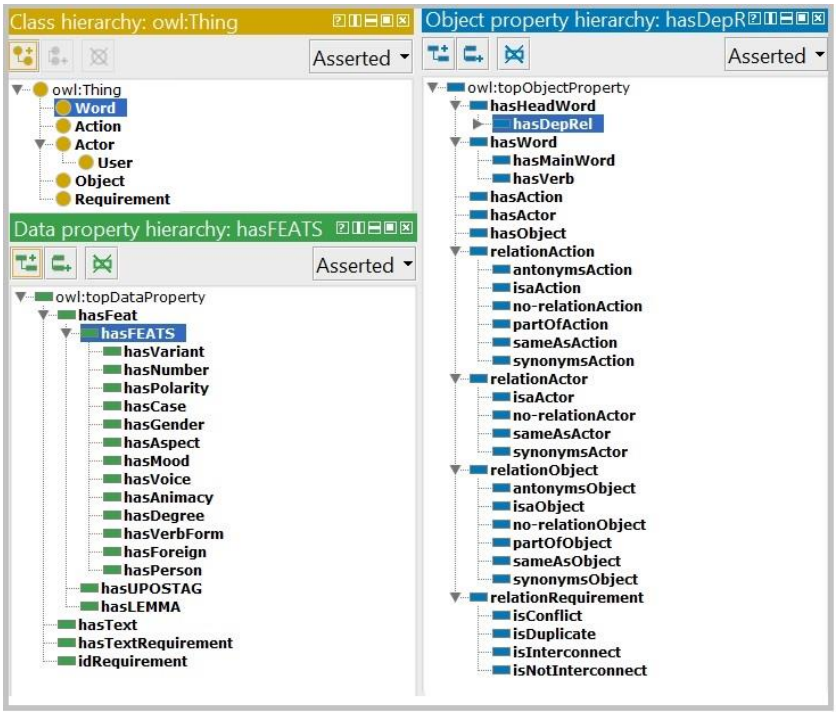


Fig. 2. Classes, object and data properties for Requirements ontology.

A set of production rules that enables to extract instances of the classes ”Actor”, ”Action”, ”Object”, ”Word” was developed to process the results of a parsing textual requirements . For instances of the class ”Word”, data properties were defined for writing tag values of speech parts and morphological features of speech parts in ac-

cordance with the UD-Russian-SynTagRus corpus model. Many developed rules can be divided according to purpose into two subsets: the rules for finding the main word in the concept of domain area and the rules for identifying all the words included in the concept. An approach to extracting requirements ontology concepts from the results of automatic processing of textual requirements is as follows:

Step 1. Processing textual requirements using the Udpipe tool

A simple text file containing a set of requirements is passed for processing to the Udpipe analyzer.

Result: dependency trees in a CoNLL-U format file.

Step 2. Extracting the actor, relation-action and object

The developed extraction rules apply to the results of the previous stage.

Result: a list of sentences indicating for each sentence the extracted names of entities and relation-actions.

Step 3. Writing data about instances in the requirements ontology

List items from the previous step are bypass. Instances of classes are added to the ontology, their data properties are indicated, object properties relations between words for the concept are set.

Result: an ontology filled with requirements instances.

3 Conclusion and future work

The proposed approach to extracting requirements ontology concepts from the results of automatic processing of requirements texts is based on the representation of knowledge about syntactic dependency tags, speech parts tags and morphological features of speech parts in the annotated Russian-language UDRussian-SynTagRus corpus. This text corpus is used by the UDPipe tool. The developed approach can be applied to other models of the Russian language that correspond to syntactic relations for Universal Dependencies. The next stage of our research will be the development of the software that will allow its user to quickly create and edit rules in design mode. As well as adding rules for searching for wording of requirements with indicating several objects to display recommendations for improving the wording of the requirements.

Acknowledgments

The research is supported by Ministry of Science and Higher Education of Russian Federation (project No. FSUN-2020-0009).

Selection of Optimal Parameters in the Fast K-Word Proximity Search Based on Multi-component Key Indexes (Extended Abstract)

Alexander B. Veretennikov[1][0000−0002−3399−1889]

Ural Federal University, 620002 Mira street, Yekaterinburg, Russia
alexander@veretennikov.ru

Abstract. Proximity full-text search is commonly implemented in modern full-text search systems. Let us assume that the search query is a list of words. It is natural to consider a document as relevant if the queried words are near each other in the document. For every occurrence of every word in a document, we employ additional indexes to store information about nearby words. We showed in previous works that these indexes can be used to improve the average query execution time by up to 130 times for queries that consist of words occurring with high-frequency. In this paper, we consider how both the search performance and the search quality depend on the values of several parameters. We propose a new index schema after the analysis of the results of the experiments.

Keywords: Full-text search · Inverted indexes · Proximity search.

1 Introduction

In full-text search, a query is a list of words. The result of the search is a list of documents containing these words. Consider a query that consists of words occurring with high-frequency. In [4], we improved the average query processing time by a factor of 130 for these queries relative to ordinary inverted indexes. The factor of proximity between the queried words in the indexed document plays an important role in modern information retrieval. Document should contain query words near each other to be relevant for the user in the context of the search query. Some words occur in texts significantly more frequently than others. An example of a typical word occurrence distribution is shown in Fig. 1. The horizontal axis represents different words in decreasing order of their occurrence in texts. On the vertical axis, we plot the number of occurrences of each word. For proximity searches, we need a word-level inverted index instead of a document-level index. The query search time is proportional to the number of occurrences of the queried words in the indexed documents. To evaluate a search query that contains words occurring with high-frequency, a search system needs much more time, as shown on the left side of Fig. 1, than a query that contains only ordinary words, as shown on the right side of Fig. 1.

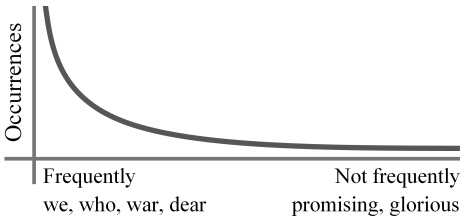


Fig. 1. Example of a word frequency distribution.

According to [2], we can consider a full-text query as a "simple inquiry". We may require that the query results be produced within two seconds, as stated in [2], to prevent the interruption of the thought continuity of the user. To improve the performance, the following approaches can be used. Early-termination approaches can be applied for full-text searches [1]. However, these methods are not effective in the case of proximity full-text searches [4]. In [3, 5], additional indexes were used to improve phrase searches, however, these approaches cannot be used for proximity full-text searches. Their area of application is limited by phrase searches. Let us discuss the stop-word approach, in which some words are excluded from the search. A word cannot be excluded from the search because even a word occurring with high-frequency can have a specific meaning in the context of a specific query [4, 5]. Therefore, excluding some words from the search can lead to search quality degradation or unpredictable effects [5]. We include information about all of the words in the indexes.

The goals and questions of this paper are as follows. We need to investigate how search performance is improved with consideration of different values of several parameters. We need to investigate search performance on the commonly used text collection. Does the search performance depend on the document size?

Our morphology analyzer provides a list of numbers of lemmas, that is, basic or canonical forms, for every word from the dictionary. Consider an array of all lemmas. Let us sort all lemmas in decreasing order of their occurrence frequency in the texts. We call the result of sorting the FL-list. The number of a lemma w in the FL-list is called its FL-number and is denoted by $FL(w)$. The first $SWCount$ most frequently occurring lemmas are stop lemmas (e.g., "time") [4]. The second $FUCount$ most frequently occurring lemmas are frequently used lemmas (e.g., "beautiful"). All other lemmas are ordinary lemmas. We use $SWCount = 500$ and $FUCount = 1050$ in the following experiments.

We use the following additional indexes [4]. The three-component key (f, s, t) index is the list of the occurrences of the lemma f for which lemmas s and t both occur in the text at distances that are less than or equal to the $MaxDistance$ from f. The two-component key (w, v) index is the list of occurrences of the lemma w for which lemma v occurs in the text at a distance that is less than or equal to the $MaxDistance$ from w. Let us consider the ordinary index with near stop word (NSW) records. For each occurrence of each ordinary or frequently used lemma in each document, we include a posting record (ID, P, NSW) in the index. ID can be the ordinal number of the specific document, and P is the position of the word in the document. The NSW record contains information

about all stop lemmas, occurring near position P in the document (at a distance that is less than or equal to the $MaxDistance$ from P). Let us discuss the results of the experiment. We create the following indexes:

$Idx0$: the ordinary inverted index without any improvements.

$Idx5$: our indexes, including the ordinary inverted index with the NSW records and the (w, v) and (f, s, t) indexes, where $MaxDistance = 5$.

$Idx7$: our indexes, where $MaxDistance = 7$.

$Idx9$: our indexes, where $MaxDistance = 9$.

GOV2 text collection and the following queries are used: title queries from TREC Robust Task 2004, title queries from TREC Terabyte Task from 2004 to 2006, title queries from TREC Web Task from 2009 to 2014, queries from TREC 2007 Million Query Track. The total size of the query set after duplicate removal is 10 665 queries. GOV2 text collection contains 25 million documents. The total size of the collection is approximately 426 GB, and after HTML tag removal, there is approximately 167 GB of plain text.

We used the following computational resources: CPU: Intel(R) Core(TM) i7 CPU 920 @ 2.67 GHz. HDD: 7200 RPM. RAM: 24 GB. All queries are evaluated within one program thread. Average query times: $Idx0$: 34.9 s, $Idx5$: 1.51 s, $Idx7$: 1.57 s, $Idx9$: 1.66 s. We improved the query time by a factor of 23.1 with $Idx5$, by a factor of 22.4 with $Idx7$, and by a factor of 21 with $Idx9$ in comparison with $Idx0$. For $Idx0$, we need 2-2.5 sec. for the search if the query consists only of ordinary or frequently used lemmas. For the queries that contain any stop lemma, $Idx0$ requires approximately 50 sec. for the search on average.

Let us consider a search query. The query consists of some set of lemmas. Let $Min\text{-}FL$-number be the minimum FL-number among all lemmas of the query. A lower FL-number corresponds to a more frequently occurring lemma. We divide the query set into subsets based on the $Min\text{-}FL$-number of queries. We select 100 as the division step. In the first subset, we include all queries with $Min\text{-}FL$-numbers from 0 to 99; in the second subset, we include all queries with $Min\text{-}FL$-numbers from 100 to 199; and so on. Let us consider the first 21.

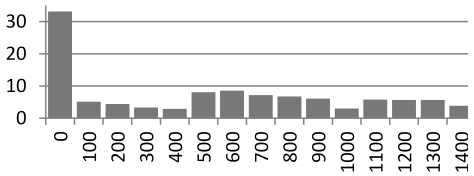


Fig. 2. The improvement factor for $Idx5$ in comparison with $Idx0$ (times); the query set is divided based on the $Min\text{-}FL$-number with a step of 100.

In Fig. 2, we show the improvement factor for $Idx5$ in comparison with $Idx0$. The first bar value is 33, which means that the average query execution time for the first subset is improved by a factor of 33. Therefore, we propose that the value of $SWCount$ can be lowered to 100. We created another index $Idx5/SW100$, with $SWCount = 100$ and $FUCount = 1450$. The results for this index look

more promising than those for $Idx5$. However, this index works relatively slow for queries that contain a stop lemma. Consequently, we propose the new index schema. The original schema can be represented by the following rules:

1) (f, s, t) indexes, where $FL(f), FL(s), FL(t) < SWCount$.

2) (w, v) indexes, where

$SWCount \le FL(w) < SWCount + FUCount$, and $SWCount \le FL(v)$.

3) Ordinary indexes (x) with NSW records, $SWCount \le FL(x)$; the NSW records contain information about all lemmas y with the condition $FL(y) < SWCount$ that occur near lemma x in the text.

For the new schema, we use the following parameters with example values. $EHFCount = 100$, $HFCount = 400$, $FUCount = 1050$.

We propose using the following indexes.

1) (f, s, t) indexes, $FL(f), FL(s), FL(t) < EHFCount + HFCount = 500$,

2) (w, v) indexes, where

$$100 = EHFCount \le FL(w) < EHFCount + HFCount + FUCount = 1450,$$
$$100 = EHFCount \le FL(v).$$

3) Ordinary indexes (x) with NSW records, $100 = EHFCount \le FL(x)$; the NSW records contain information about all lemmas y with the condition $FL(y) < EHFCount = 100$ that occur near lemma x in the text.

We investigated how multi-component key indexes help to improve search performance. We used well-known GOV2 text collection. We proposed a new index schema. We analyzed how the value of $MaxDistance$ affects the search performance. With an increase in the value of $MaxDistance$ from 5 to 9, the average search time using multi-component key indexes was increased from 1.51 sec. to 1.66 sec. and the value of $MaxDistance$ can be increased even further.

References

1. Anh, V.N., de Kretser, O., Moffat, A.: Vector-space ranking with effective early termination. In: Proceedings of the 24th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval. pp. 35–42. SIGIR '01, ACM, New York, NY, USA (2001). https://doi.org/10.1145/383952.383957
2. Miller, R.B.: Response time in man-computer conversational transactions. In: Proceedings of the December 9-11, 1968, Fall Joint Computer Conference, Part I. pp. 267–277. AFIPS 68 (Fall, part I), ACM, USA (1968). https://doi.org/10.1145/1476589.1476628
3. Petri, M., Moffat, A.: On the cost of phrase-based ranking. In: Proceedings of the 38th International ACM SIGIR Conference on Research and Development in Information Retrieval. p. 931934. SIGIR 15, Association for Computing Machinery, New York, NY, USA (2015). https://doi.org/10.1145/2766462.2767769
4. Veretennikov, A.B.: Proximity full-text search by means of additional indexes with multi-component keys: In pursuit of optimal performance. In: Manolopoulos Y., Stupnikov S. (eds) Data Analytics and Management in Data Intensive Domains. DAMDID/RCDL 2018. Communications in Computer and Information Science. vol. 1003, pp. 111–130. Springer. https://doi.org/10.1007/978-3-030-23584-0_7
5. Williams, H.E., Zobel, J., Bahle, D.: Fast phrase querying with combined indexes. ACM Trans. Inf. Syst. **22**(4), 573–594 (Oct 2004). https://doi.org/10.1145/1028099.1028102

Data Driven Detection of Technological Trajectories[1] (Extended Abstract)

Sergey Volkov[I,II], Dmitry Devyatkin [I], Ilya Tikhomirov[I], Ilya Sochenkov[I]

[I]Federal Research Center "Computer Science and Control" RAS, Moscow, Russia
[II]Peoples' Friendship University of Russia, Moscow, Russia

Abstract. The paper presents a text mining approach to identifying and analyzing technological trajectories. The main problem addressed is the selection of documents related to a particular technology. The approach includes new keyword and keyphrase detection method, word2vec embeddings-based similar document search method and fuzzy logic-based methodology for revealing technology dynamic. The experimental database contains more than 4.7 million documents. Self-driving car technology was chosen as an example. The result of the experiment shows that the developed methods are useful for effective searching and analyzing information about given technologies.

Keywords. Technological trajectories, text mining, keywords, similar documents retrieval.

1 Introduction

"Technological paradigm" is a specific set of technological innovations, which has its list of "relevant" problems as well as the procedures and the specific knowledge to solve these problems [1]. The development path within a technological paradigm is a technological trajectory. In the paper, we propose a text mining approach to identifying and analyzing technological trajectories. We propose a new keyword and keyphrase detection method to deal with the first issue, which combines the topic importance characteristic and neology score of terms. We then utilize the word embedding-based method to identify a technological trajectory [2] and apply the new keyword score to make the search more technology-focused. In the analyzing part, we mainly consider the conditions which precede a breakthrough. To overcome the second issue, we created a fuzzy logic-based method for revealing technology dynamics.

2 Description of the dataset and methods

The United States Patent and Trademark Office (USPTO) database is the primary data source for the experiments [4]. First, we have applied TextAppliance web crawling

[1] This study was supported by Russian Foundation for Basic Research, grant number 17-29-07016 ofi_m

tools [5] to download those web-pages, and then chose the technology for the test, "Self-driving car". After that, we manually selected 10 "seed" patents with the integrated USPTO search system. Then we assembled a gold dataset to evaluate the accuracy of the presented methods with the following steps. First, we built a list that combines search results of all the considered methods with the results obtained by human analytics with the USPTO search tool. All 4.7 million documents were used to perform the searches. Second, we submitted that list of documents to human annotators who excluded all non-relevant documents from the list. Finally, all the remaining patents are related to the self-driving car technology or the technologies that could have preceded it [6].

The method for an identification of a technology contains the following steps. The first step is keyword and key-phrase extraction for the seed documents. In this case keywords are the most important lexis of a document [7]. We applied well-known *tf-idf* score to assess that importance. After that the document dates are used to re-weight the keywords. For that we utilize a novelty score which is a frequency of using of some word in a particular date interval. In this study 5 date intervals from 1996 to 2020 were considered; therefore we obtain particular value for an each interval. Third step is sorting keyword and keyphrases with the obtained weights. We obtain top-50 keywords for each pair <*interval, document*> and use them in the following steps. Forth step includes training Word2Vec models on documents, related to the analyzed intervals. We got 5 different models as the result. The next step is seeking documents that are similar to the previously obtained ones.

1. Vectorize all keywords from the current document list with the model for the current interval (begin with «2013-2020»).
2. Sum the obtained vectors for each document and form the result vector. This way each document is represented with one vector.
3. Search for similar documents in the whole patent collection. We use cosine similarity for evaluating the distance between documents.
4. Extend the current document list by choosing top-n of the most similar document from the obtained in step 3.
5. Repeat all steps for the earlier date interval.

Method for technology dynamic revealing is based on the analysis of per-year cumulative patent count for the identified technologies. We applied logistic curve (S-curve) [8] for the approximation of the cumulative patent count. $f: \mathbb{R} \to \mathbb{R}^+$, $f(x) = \frac{C}{b+e^{-ax}}$., where $a, b, C \epsilon \mathbb{R}^+$, – parameters of that function are obtained with the least-squares approach, and x is a normed year. The applicability of that curve for technology growth modeling has been proved, for example in [9].

The cumulative patent count has the highest speed of growth if the first derivative of the S-curve function reaches the maximum. We first find the argument x_{max}, for that maximum: $x_{max} = -\ln(b)/a$. Define a linguistic variable "Technology development dynamics", which can take values from the set $X = \{X_h, X_l, X_m\}$, there $X_h = 'high'$, $X_l = 'low$, $X_m = 'middle'$.

Define a fuzzy set for each value of the linguistic variable. Those fuzzy sets contain arguments of $f(x)$. Let define membership functions for that: $\mu_{h,m,l}: \mathbb{R} \to [0,1]$.

First of all, define a membership function for the set which is related to X_h. As said earlier speed of patent growth depends on $f'(x)$. That derivative is always positive and upper-bounded. Accordingly to the condition $\mu \leq 1$ we find it uniform norm:

$$\mu_h(x) = 4bf'(x)/\mathrm{Ca}$$

Hence value "low" is semantically negative to "high", one can define the membership function for that value as follows: $\mu_l(x) = 1 - \mu_h(x)$. Value "middle" describes situation, there $\mu_l(\mathrm{x}) = \mu_h(\mathrm{x})$, therefore the membership is the following: $\mu_m = 1 - |\mu_h(x) - \mu_l(x)|$.

Finally, if one wants to evaluate the development dynamic of technology at some year *x*, they should choose the membership function which returns the maximum and then find the value of the linguistic variable for that function. Accordingly to [3], we assess the state of the analyzed technology as well as states for the related technology on particular dates and try to find a relationship between them.

3 Results

First method is sequential search. There are five steps corresponding to five time periods. On each step similar documents retrieval applies only to the documents, which were found on the previous step. Four cases were considered (for top 5,10,20,50 keywords which are sorted by new coefficient). On each step 50 documents were found. The document is considered valid if it is contained in the experts-filtered gold dataset.

Table 1. *Accuracy of the search*

Search type	Step/ Top-N Keywords	1 (2013-2020)	2 (2009-2014)	3 (2005-2010)	4 (2000-2006)	5 (1996-2001)	avg
Sequential search	5	**0.9**	0.66	0.68	0.7	0.26	0.64
	10	**0.9**	0.2	0.16	0.62	0.02	0.38
	20	0.88	0.46	0.56	0.86	**0.96**	0.74
	50	**0.9**	**0.96**	**0.8**	**0.94**	0.92	**0.904**
Full search	5	**0.9**	0.66	0.66	0.14	0.12	0.5
	10	**0.9**	0.2	0.32	0.48	0.02	0.38
	20	0.88	0.46	0.56	0.86	**1**	0.75
	50	**0.9**	**0.96**	**0.8**	**0.9**	0.98	**0.908**

The second method is the full search. It differs from the first one in that similar documents retrieval is processed at each step not only for patents, found on the previous step, but for all current documents. Table 1 shows step-by-step accuracy for the both methods. To assess the quality of our method, we compared it with the pre-trained BERT model (bert-base-cased). Although BERT performs slightly better in the initial

steps, it shows lower accuracy in total (0.908 – our model avg. accuracy, 0.852 – BERT avg. accuracy). We believe this is because the novelty score helped to filter non-technical lexis, which allows the proposed method to stay on the track.

Finally, we analyzed the cumulative patent dynamic for the self-driving-related technologies and built fuzzy-set membership functions for them. The results showed that the breakthrough in the autonomous driving field is preceded with high development dynamics of the related technologies, namely the highest dynamics was for "car computer vision", "traffic control", "navigation".

4 Conclusion

The experiment results show that the developed methods are useful for technology-related document retrieval and breakthrough-technologies prediction. Unfortunately, it is not tricky to detect a technological trajectory using a single source. In the future, we are going to increase the number of information sources and collect extensive statistics on the development of various technologies. Namely, we will consider not only patents but also scientific papers, theses, technical reports, and other sources.

The detailed analysis of obtained patents shows how the technologies develop in time and predict future development directions. In addition to a breakthrough prediction, we are going to consider additional technology-state related events, such as the emergence of new technology.

References

1. Dosi, G.: Technological paradigms and technological trajectories: a suggested interpretation of the determinants and directions of technical change. Res. Policy 11(3), pp.147–162 (1982).
2. Volkov S. S. et al. Towards Automated Identification of Technological Trajectories //Russian Conference on Artificial Intelligence. – pp. 143-153. Springer, Cham (2019).
3. Napolitano L. et al. Technology networks: the autocatalytic origins of innovation //Royal Society open science. Vol. 5(6), pp. 172445 (2018).
4. Search for patents–USPTO. https://www.uspto.gov/patents-application-process/search-patents. Last accessed: 2020/07/30.
5. Osipov G. et al. Exactus expert—search and analytical engine for research and development support //Novel Applications of Intelligent Systems. pp. 269-285. Springer, Cham (2016).
6. Dataset trajectories-uspto-v2. http://nlp.isa.ru/trajectories-uspto-v2. Last accessed 2020/05/19.
7. Suvorov R. E., Sochenkov I. V. Establishing the similarity of scientific and technical documents based on thematic significance //Scientific and Technical Information Processing. Vol. 42(5), pp. 321-327 (2015).
8. Foster R. N. Working the S-curve: assessing technological threats //Research Management. Vol. 29(4), pp. 17-20 (1986).
9. Andersen B. The hunt for S-shaped growth paths in technological innovation: a patent study //Journal of evolutionary economics. Vol. 9(4), pp. 487-526 (1999).

Comparison of cross-lingual similar documents retrieval methods * (Extended Abstract)

D.V. Zubarev[1][0000−0002−9687−6650] and I.V. Sochenkov[1][0000−0003−3113−3765]

Federal Research Center 'Computer Science and Control' of Russian Academy of Sciences, 44-2 Vavilov Str., Moscow 119333, Russia. zubarev@isa.ru

Abstract. In this paper, we compare different methods for cross-lingual similar document retrieval for distant language pair, namely Russian and English languages. We compare various methods among them: classical Cross-Lingual Explicit Semantic Analysis (CL-ESA), machine translation methods and approaches based on cross-lingual embeddings. We introduce two datasets for evaluation of this task: Russian-English aligned Wikipedia articles and automatically translated Paraplag. Conducted experiments show that an approach with inverted index, with an extra step of mapping top keywords from one language to other with the help of cross-lingual word embeddings, achieves better performance in terms of recall and MAP than other methods on both datasets.

Keywords: cross-lingual document retrieval · cross-lingual plagiarism detection · cross-lingual word embeddings.

1 Introduction

This paper is a continuation of our previous study [5]. Document retrieval from a big collection of texts is important information retrieval problem. This problem is extensively studied for short queries, such as user queries to search engines. The document retrieval with texts as queries imposes some difficulties, among them an inability to capture the main ideas and topics from the long text. The problem becomes even harder when we enter the field of cross-lingual document retrieval. Some tasks require to use a text (possibly long) as a query to retrieve documents that are somehow similar to it. One of these tasks is plagiarism detection.

2 Document retrieval methods

In this section, we describe various methods that we used for similar document retrieval.

* The reported study was funded by RFBR according to the research projects No 18-37-20017 & No 18-29-03187.

2.1 Preprocessing

On a preprocessing stage, we split each sentence into tokens, lemmatize tokens and parse texts. We use Udpipe [1] for the Russian language and for the English language. Besides, we removed words with non-important part of speech: conjunction, pronoun, preposition, etc., and common stop-words (be, the, etc.).

2.2 Cross-lingual embeddings

We train cross-lingual word embeddings for a Russian-English pair on parallel sentences available on the Opus site. All parallel sentences are preprocessed. After that, all pairs that have a difference in the size of more than 10 words are filtered out. We use syntactic phrases up to 3 words in length to enrich the vocabulary.

We apply a method proposed in [4], designed for learning bilingual word embeddings from a non-parallel document-aligned corpus, but it can be used for learning on parallel sentences too. Words are inserted into the pseudo-bilingual sentence relying on the order in which they appear in their monolingual sentences and based on a length ratio of two sentences. After that, the word2vec skip-gram model is used on the resulting bilingual corpus.

2.3 Retrieval-Based Approach

We use a custom implementation of inverted index [3], which maps each word to a list of documents in which it appears along with weight that represents the strength of association of this word with a document. Along with words, we index syntactic phrases up to 3 words. At query time, we extract the top terms from the query document according to some weighting scheme. Then we map each keyword to N other language keywords with cross-lingual embeddings. Then we use either classical ranking function like BM25 or calculate similarity score between the query vector and all other vectors (terms are weighted with TF*IDF).

2.4 Translation method

This method resembles a translation method that was introduced in [2]. Basically, this method is the same as the previous one, with the exception that related terms are merged into the term, to which they are related, after the retrieval stage. The weight of each term is adjusted according to the formula:

$$\widehat{w}_t = w_t + \sum_{t' \in R(t)} P_T(t|t') w_{t'} \tag{1}$$

where $R(t)$ is a set of related terms, $P_T(t|t')$ is the translation probability. In this case, translation probability is a similarity score.

[1] russian-syntagrus-ud-2.5-191206 and english-ewt-ud-2.5-191206 models

2.5 Machine translation

It is quite a natural approach: translate the query text to the other language and perform monolingual retrieval of similar documents. For training machine translation model, we used a subset of the same parallel sentences as for training cross-lingual embeddings.We used OpenNMT-py library with default settings to train a machine translation model.

2.6 Document as vector

In this approach, we represent each document as a dense vector. It is done by averaging vectors of the top K keywords of the document. After that, we index all vectors with ANN index.

2.7 Sentence as vector

This method is similar to the previous one. Words with low IDF (<0.01) are removed from sentences. After that, if the length of the sentence is greater or equal than 3, we average sentence terms' embeddings to obtain sentence embeddings. Those vectors also indexed with ANN index. At search time, we search for the m most similar sentence vectors for each query sentence. After that the documents are scored based on similarity and number of retrieved sentences.

2.8 Explicit semantic analysis (ESA)

We implemented CL-ESA method described in [1]. This method represents the document as a weighted vector of concepts. We selected around 800 000 English articles that are aligned with Russian Wikipedia articles. We precomputed vector of concepts for each document in text collection and indexed them. At query time, the query document is converted to a vector of weighted concepts, i.e., identificators of Wikipedia articles. Then those identificators are mapped to articles in the other language, and similar documents are retrieved via search in the inverted index.

2.9 Clustering of word embeddings

This approach is similar to ESA method, but instead of using concepts from Wikipedia, it uses centroids of clusterized embeddings. We used K-means method for clustering cross-lingual word embeddings into 100k clusters. We take top K keywords of the document and search n most similar centroids for each keyword. After that, we sum similarities to centroids to obtain a weighted vector of centroids for each document.

3 Evaluation Results

As in our previous work, we use Russian-English aligned Wikipedia articles as a dataset for evaluation of retrieval methods. Also, we translated the text of sources from Paraplag dataset from Russian to the English language. All sources from two datasets were combined into one collection. The most similar 150 documents were retrieved and evaluated by standard metrics: Recall, MAP.

Table 1 displays the evaluation results.

Table 1. Evaluation results.

	Essays				**Wiki**			
Method	**P@1**	**Rec@10**	**Rec**	**MAP**	**P@1**	**Rec@10**	**Rec**	**MAP**
TM, bm25	**0.98**	**0.917**	**0.985**	**0.896**	**0.503**	**0.732**	**0.829**	**0.584**
RBA, bm25	0.98	0.905	0.983	0.888	0.508	0.717	0.82	0.584
MT, bm25	0.793	0.758	0.949	0.687	0.358	0.49	0.675	0.404
DocVec	0.473	0.347	0.675	0.274	0.283	0.478	0.672	0.350
SentVec	0.306	0.338	0.642	0.251	0.292	0.525	0.685	0.367
Cent	0.273	0.197	0.395	0.151	0.203	0.29	0.58	0.25
ESA	0.254	0.453	0.802	0.218	0.183	0.35	0.602	0.28

The results show that the translation method with BM25 ranking function is better in terms of Recall and Map than other methods on both datasets. The RBA with BM25 is slightly worse than TM. A simple method based on machine translation can't compete with the methods based on embeddings. The performance of vector based methods is ruined when searching over a large number of source vectors.

References

1. Potthast, M., Barrón-Cedeño, A., Stein, B., Rosso, P.: Cross-language plagiarism detection. Language Resources and Evaluation **45**(1), 45–62 (2011)
2. Rekabsaz, N., Lupu, M., Hanbury, A., Zuccon, G.: Generalizing translation models in the probabilistic relevance framework. In: Proceedings of the 25th ACM international on conference on information and knowledge management. pp. 711–720 (2016)
3. Sochenkov, I.V., Zubarev, D.V., Tikhomirov, I.A.: Exploratory patent search. Informatika i Ee Primeneniya [Informatics and its Applications] **12**(1), 89–94 (2018)
4. Vulic, I., Moens, M.F.: Bilingual word embeddings from non-parallel document-aligned data applied to bilingual lexicon induction. In: Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics (ACL 2015). vol. 2, pp. 719–725. ACL; East Stroudsburg, PA (2015)
5. Zubarev, D.V., Sochenkov, I.V.: Cross-lingual similar document retrieval methods. Proc. ISP RAS **31**(5) (2019)

On Developing of the FrameNet-like Resource for Tatar (Extended Abstract)

Ayrat Gatiatullin[1,2], Alexander Kirillovich[1] and Olga Nevzorova[1]

[1] Kazan Federal University, Kazan, Russia
[2] Tatarstan Academy of Sciences, Kazan, Russia
ayrat.gatiatullin@gmail.com, alik.kirillovich@gmail.com, onevzoro@gmail.com

In this paper, we present TatVerbBank, the first FrameNet-like resource for Tatar language. This project is inspired by FrameNet and FrameBank [1].

Russian FrameBank is a bank of annotated samples with lexical constructions (e.g. argument constructions of verbs and nouns) from the Russian National Corpus. FrameBank belongs to FrameNet-oriented resources, but unlike Berkeley FrameNet it focuses more on morphosyntactic and semantic features of individual lexemes rather than on generalized frames.

In FrameNet the central element is the frame, but in FrameBank the lexeme is the central element and individual lexeme has its own set of lexical constructions.

FrameBank serves as a main prototype for developing the Tatar VerbBank resource (hereinafter TatVerbBank). TatVerbBank is a FrameBank-like resource which consists of a set of annotated samples with verb constructions from the "Tugan tel" Tatar National Corpus (http://tugantel.turklang.tatar).

Our goal is to create a dictionary of verb constructions that would contain semantic and especially syntactic information about verb actants in valency grammar. We use a reduced model of construction descriptions in Tatar compared with FrameBank, which includes:

1. the syntactic rank of the element (Subject, Object, Predicate, Peripheral, Clause);
2. the morphosyntactic features of the element (including POS, affix marking);
3. the semantic roles of the argument (e.g., Agent, Patient, Instrument);
4. the lexical-semantic class of the element (e.g., human, animate, abstract entity, means of transport, etc.);
5. one or several examples.

A base hierarchy of predicates and semantic roles is defined in FrameBank. The detailed list of semantic roles in this resource currently contains 91 items classified into seven domains such as Agent, Possessives, Patient, Addressee, Experiencer, Instrument and Settings. TatVerbBank uses the same set of semantic roles that are defined in FrameBank, so the selected main roles uniquely define the frame and peripheral roles are used for event aspects in general.

When building TatVerbBank, we are using various lexical resources for the Tatar language. The main lexical resource is the Russian-Tatar explanatory dictionary by Ganiev F.A. Another lexical resource is the Russian-Tatar Social-Political thesaurus

depeloped by Institute of Applied Semiotics of Tatarstan Academy of Sciences. At the first stage, we developed the verb dictionary which consists of Tatar lexemes denoting events, phenomena or processes. Then we grouped words into "sense groups" and built a basic structure (basic frame) for each group. The verbs (concepts) from the verb dictionary can be ambiguous and have different senses. We then linked this dictionary to the Russian-Tatar socio-political thesaurus in order to reveal the existing ambiguity using the systems of concepts of these resources. Thus we got different senses of ambiguous words.

The TatVerbBank resource has a multi-level structure in which verb constructions (we called them situational frames) are included in a complex structure with the dictionary of verb constructions and the Russian-Tatar thesaurus. Each concept of the Russian-Tatar thesaurus has lexical inputs and each basic frame has a set of instances.

This model is the closest to the general data model in FrameBank; here the verb construction is built into the frame as a set of grammar restrictions applied to frame slots. The frame model used to describe the construction of a verb contains basic frames with slots which defined as semantic roles and instances of frames with grammatical and syntactic restrictions of slot values.

Instances of frames were taken from the "Tugan Tel" electronic corpus of the Tatar language (http://tugantel.tatar/).

The TatFrameBank resource is published at Linguistic Linked Open Data cloud [5]. We used Lemon [2], LexInfo [3] and PREMON [4] ontologies for representing data.

References

1. Lyashevskaya, O., Kashkin, E.: FrameBank: A Database of Russian Lexical Constructions. In: Khachay, M., et al (eds). Proceedings of the 4th International Conference on Analysis of Images, Social Networks and Texts (AIST 2015). Communications in Computer and Information Science, vol 542, pp. 350-360. Springer (2015). doi:10.1007/978-3-319-26123-2_34
2. McCrae, J.P., Bosque-Gil, J., Gracia, J., Buitelaar, P., and Cimiano, P.: The OntoLex-Lemon Model: Development and Applications. In: Kosem I., et al. (eds.) Proceedings of the 5th biennial conference on Electronic Lexicography (eLex 2017), pp. 587–597. Lexical Computing CZ (2017).
3. Chiarcos, C.: OLiA – Ontologies of Linguistic Annotation. Semantic Web **6**(4), 379–386 (2015). https://doi.org/10.3233/SW-140167.
4. Rospocher, M., Corcoglioniti, F., and Palmero Aprosio, A.: PreMOn: LODifing linguistic predicate models. Language Resources and Evaluation **53**, 499–524 (2019). https://doi.org/10.1007/s10579-018-9437-8.
5. Cimiano, P., Chiarcos, C., McCrae, J.P., and Gracia, J.: Linguistic Linked Open Data Cloud. In: Cimiano, P., et al. (eds.) Linguistic Linked Data: Representation, Generation and Applications, pp. 29–41. Springer (2020). https://doi.org/10.1007/978-3-030-30225-2_3.

PHD WORKSHOP

Взаимное отображение графовой и реляционной моделей данных для мультимодельных баз данных (Расширенные тезисы)

Аркадий Ошеев[1]

[1] Московский государственный университет им. М.В. Ломоносова, Москва, Россия
arkadii_osheev@mail.ru

Современные информационные системы одновременно используют данные различной структуры. Для их представления уже недостаточно реляционной модели данных. Также, концептуальные схемы многих предметных областей значительно удобнее представлять в NoSQL моделях, например, в графовой модели. Использование в рамках одной информационной системы нескольких СУБД, реализующих разные модели данных, приводит к существенному усложнению управления системой.

Одним из возможных решений данной проблемы является использование мультимодельной СУБД. Формальной основой для разработки такой СУБД является взаимное отображение между применяемыми моделями данных.

Эта работа посвящена теме интеграции реляционной и графовой моделей. Реляционная модель является самой популярной в мире и занимает 60% рынка на 2019 год[1]. Графовая модель естественна для представления некоторых предметных областей, например, социальных сетей. В качестве языка запросов для графовой модели берется Cypher, для реляционной модели — SQL.

В настоящее время развиваются мультимодельные СУБД, поддерживающие графовую и реляционную модели такие как Agens Graph, CosmosDB, OrientDB, MarkLogic.

Несмотря на наличие данных СУБД, вопрос интеграции реляционной и графовой моделей не решен полностью. Одной из проблем является разнообразие языков манипулирования графовыми данными: Cypher, SPARQL, Gremlin, PGQL.

Для интеграции интересующих моделей MarkLogic и OrientDB отображают данные в промежуточное представление. Agens Graph не предоставляют отображение моделей друг в друга, а просто хранит данные графовой и реляционной моделей в раздельных хранилищах и для одновременного обращения к этим хранилищам необходимо производить гибридные запросы.

Таким образом, по сведениям автора, в настоящее отсутствуют полноценные подходы к взаимному отображению реляционной и графовой моделей без использования промежуточного представления.

---

[1] https://scalegrid.io/blog/2019-database-trends-sql-vs-nosql-top-databases-single-vs-multiple-database-use/

Целью данной работы является разработка взаимного отображения графовой и реляционных моделей для обеспечения возможности обращения на реляционном языке запросов к графовым данным и наоборот. Такое отображение может быть применено в качестве формальной основы для мультимодельной СУБД.

Для достижения данной цели необходимо решить следующие основные задачи:

- Произвести обзор существующих мультимодельных баз данных, реализующих графовую и реляционную модели.
- Определить варианты взаимного отображения схем графовой и реляционной моделей.
- Определить варианты взаимного отображения языков запросов реляционной и графовой моделей.
- Реализация отображений схем и запросов для конкретных выбранных реляционной и графовой СУБД.
- Произвести сравнение эффективности вариантов отображения на тестовых базах данных и наборах запросов.

В данной работе был произведен обзор существующих мультимодельных СУБД, предложено по одному варианту основных идей отображений (i) реляционной схемы в графовую схему, (ii) графовой схемы в реляционную схему, (iii) запросов Cypher в SQL, (iv) запросов SQL в Cypher.

Обзор затрагивает следующие мультимодельные СУБД, объединяющие интересующие модели: Agens Graph[3], CosmosDB[4], OrientDB[5], MarkLogic[6]. Сравнение производилось на основании документации по следующим критериям: поддерживаемые языки запросов, модели данных и некоторые аспекты интеграции данных; сводная информация представлена в виде таблицы в статье, на ее основе были сделаны выводы, показывающие отсутствие полноценного отображения между графовой и реляционной моделями.

Источником идей для отображения схем графовой и реляционной моделей являются работы [4][5]. В работе [4] предложен алгоритм трансформации реляционных данных в графовое представление. Представлены идеи, как при обходе отношений выяснить, что является вершиной, а что ребром. В работе [5] изложены рекомендации Neo4j по переходу от реляционного представления в графовое. Результатом анализа вышеперечисленных работ являются идеи по отображению реляционной и графовой моделей друг в друга. Данные идеи снабжены примерами и изложены в тексте статьи.

Основанием для идей отображений между языками запросов SQL и Cypher являются работы [1][2][3]. В работе [1] приводится формальное описание Cypher и его особенности. Работа [2] посвящена формализации запросов на языке Cypher посредством адаптации операторов реляционной алгебры для работы с графами, также авторы представляют новые операторы. В работе [3] представ-

[3] https://bitnine.net/documentations/manual/agens_graph_developer_manual_en.html
[4] https://docs.microsoft.com/ru-ru/azure/cosmos-db/
[5] http://www.orientdb.com/docs/last/index.html
[6] https://docs.marklogic.com

лен инструмент, который теоретически мог отображать Cypher в любой язык запросов, при правильном формировании файла топологического отображения gTop. Идеи по отображению языков запросов представлены в статье и снабжены примерами.

Несмотря на проделанную работу, в данной статье не учтены важные детали рассматриваемых моделей. Например, при отображении реляционной схемы в графовую учитывается только случай, когда внешний ключ состоит из одного атрибута. Этот и другие детали будут рассматриваться в дальнейшей работе.

На данный момент основными направлениями дальнейшей работы являются:

- учет при отображении важных понятий реляционной и графовой моделей, таких, как функциональные зависимости различного вида, составные внешние ключи и т.д.;
- исследование других вариантов отображения между схемами и языками запросов (например, отображение каждого типа отношения в отдельную вершину графа). Планируется рассмотреть влияние функциональных зависимостей на отображение;
- формализация алгоритмов отображения схем и запросов между Cypher и SQL. Языки запросов планируется формализовать с использованием реляционной алгебры и реляционной алгебры над графами, предложенной в работе[2]. Для формализации правил отображения языков запросов планируется использовать движимый моделями подход (MDA) и язык отображения моделей ATL[7];
- реализация отображений схем и запросов для конкретных выбранных реляционной и графовой СУБД;
- сравнение эффективности вариантов отображения на тестовых базах данных и наборах запросов.

Литература

1. Francis, Nadime & Taylor, Andrés & Green, Alastair & Guagliardo, Paolo & Libkin, Leonid & Lindaaker, Tobias & Marsault, Victor & Plantikow, Stefan & Rydberg, Mats & Selmer, Petra. (2018). Cypher: An Evolving Query Language for Property Graphs. 1433-1445. 10.1145/3183713.3190657.
2. Marton, József & Szárnyas, Gábor & Varro, Daniel. (2017). Formalising openCypher Graph Queries in Relational Algebra. 182-196. 10.1007/978-3-319-66917-5_13.
3. Steer, Benjamin & Alnaimi, Alhamza & Lotz, Marco & Cuadrado, Félix & Vaquero, Luis & Varvenne, Joan. (2017). Cytosm: Declarative Property Graph Queries Without Data Migration. 1-6. 10.1145/3078447.3078451.
4. Virgilio, Roberto & Maccioni, Antonio & Torlone, Riccardo. (2013). Converting relational to graph databases. 10.1145/2484425.2484426.
5. https://neo4j.com/developer/guide-importing-data-and-etl/

[7] https://www.eclipse.org/atl/

Towards ontology-based cyber threat response (Extended Abstract)

Nikolay Kalinin

Faculty of Computational Mathematics and Cybernetics, MSU

An area of information security today is especially relevant: the number of threats and their destructive capacity grows with every year. In such circumstances, the development of new methods that would be applied in composition with the automated tools of exposure of cyber threats becomes not simply an interesting scientific task, but also a valuable practical result.

Two main approaches to building such systems - signature approach and machine learning methods have several weaknesses: bases of signature need continuous support, Machine learning algorithms is difficult to interpret and configure. To two indicated approaches we can add an approach on the basis of formal models. As noted in the work [9] ontology is already one of the fixed assets of realization of the large systems of information security. Tools of exposure threat on the basis of formal models can allow not only to identify and classify threats but also to effectively produce reliable and interpreted decisions for their removal. he purpose of this work is a demonstration of wide possibilities of an ontological approach for the development of methods and tools for reacting to the security threats of the distributed information infrastructure.

The construction of ontological models in information security is conducted already for more than fifteen years. One of the first bright works in this area is ontology IDS [1], presented in work [11]. Authors put the aim to show the utility of ontology as a model for classification of attacks in the intrusion detection system underlining their superiority above more used taxonomies. Their result ontology presented as an attack classification framework and described in DAML-OIL [2] (language predecessor OWL[4]).

Approach allowing to systematize not only information security but also the development of ontology process, presented in works [5] and [6]. In the first authors examine the methodology of construction of ontology in cybersecurity. In the authors opinion, ontologies are usually an association of three levels, from most general, such as DOLCE, at the top level, to the applied ontology [6], this approach gets further development described as full ontology-framework CARTELO. The ontology DOLCE- SPRAY is used at the top level, at middle-level ontology is presented by the ontology of SECCO, plugging in itself the basic concepts of cybersecurity, the ontology of cyber-operations OSCO complements at the bottom level.

In [3] authors note that formal representation of knowledge and integration of information from different sources allows substantially improve quality of exposure and response. In the article as main scenarios of the use are the search of

[1] IDS - Intrusion detection system

relevant records from IDS, collection of information about software, and attempt of determination of malicious activity on the basis of network traffic and changes in the system. For the solution of these tasks, the authors develop the ontology of STUCO. Its notable features are relative simplicity and realization by means of the JSON- scheme.

The common decision of long-term problem standardization of formats of cybersecurity-related knowledge lately became language STIX [1], therefore no wonder that the most complex is the universal ontology of cybersecurity (UCO), presented in work [10] is based on exactly its structure. An offered ontology is implemented in OWL DL assuming an effective inference allows us to extract information from all popular industrial dictionaries and assumes the wide spectrum of scenarios of the use.

In the conclusion of this review, we want to note that for the past years substantial results were obtained many tasks are not solved. Possibility of reasoning is not used, the problem of extraction of knowledge from the unstructured sources is not fully resolved, ontologies do not contain concepts for the description of information infrastructure. A possible way for efficient infrastructure representation presented in recent work [7], but valuable ontologies containing both knowledges about infrastructure and knowledge of information security are yet to be developed.

To show the possibilities of the ontological approach the model knowledge base was implemented. A terminological constituent (T - box) of the knowledge base is the ontology of UCO complemented applied with bottom level ontologies. An actual constituent (A - Box) plugs operating information (events and incidents of cybersecurity), information about infrastructure, and also information from open dictionaries and taxonomies.

As an adaptation for UCO, additional ontologies for the decision of certain tasks were developed. The ontology of operating information extends and complements such concepts of UCO as action and observable. Its main task is to provide accordance with other objects of knowledge base and by operating information. Ontology of information infrastructure is a clarification for uco identity - local identity that allows to determine authentication for internal subjects and also essence set of infrastructure objects for a description of endpoints in an infrastructure. Last from ontology models is prioritization ontology. It is a model for a conclusion of environmental risk metrics CVSS. It includes concepts from the environmental risk of CVSS.

In a model knowledge base operating information appears as events of SIEM. SIEM events are the records of compatible format, extracted from different logs and security tools aggregated in a single database. The format of records is based on the Cybox standard (http://cybox.mitre.org), plugged into STIX. As a model data, the logs of regular subsystem of audit of OS Linux, logs generated by Osquery framework (https://osquery.io), and logs of firewalls were used.

Information about an infrastructure appears in two basic types of objects: endpoint record and network rule. The first type contains information about a certain host, the second is an object for network availability description and

written down like the rules of firewalls.The main entities for describing endpoint record are port, interface (IP), and route. The main entities associated with the Network rule are the source and destination IP subnets (or a single instance of the interface) and the type of rule (access control rule or address translation rule). The resulting records allow us to use the mechanism of logical reasoning to build the reachability matrix as show at [7].

For the developed model, three main use cases can be distinguished: classification of attacks, assessment of the risk level, and finding related information.

Attack classification is possible by using logical reasoning because it allows clarifying abstract class (SIEM event in our model) to the concrete attack class. This concretizing based on properties and links of an instance of the object. In particular, such clarify allow to avoid errors of a single sensor.

The risk level is estimated in accordance with the second version CVSS standard. The standard of CVSS is plugged in itself by three types of metrics: base, temporal, and environmental. The first two metrics are descriptions of vulnerability presented in the ontology. For the calculation of environmental metrics descriptions of the affected objects are used.

The finding of related information in our model can be materialized on the basis of rules presenting as SPARQL [8] queries. Such SPARQL queries must be certain for every type of event at the level of the user interface.

Within the work, the model of knowledge base was built to support processes of response to the threats of information security. The special attention at the development of the ontological model was spared to a description of information infrastructure. But there is still a long way to go for using this model. Firstly, in work, we did not involve the question of possibility of thread data processing and, as a result, the productivity questions. Secondly, a fairly primitive model for describing network availability was used. Thirdly, valuable use of the system is impossible without serious expansion of types of processed events and expansion of the set of concepts in ontologies of the application layer. Our global aim is to develop a complete ontological framework for support of the response to cyber threats and this research is only the first step on a path to this aim.

This work is supervised by Nikolay Skvortsov, Federal Research Center Computer Science and Control of the Russian Academy of Sciences (FRC CSC RAS).

References

1. Barnum, S.: Standardizing cyber threat intelligence information with the structured threat information expression (stix). Mitre Corporation **11**, 1–22 (2012)
2. Horrocks, I., et al.: Daml+oil: A description logic for the semantic web. IEEE Data Eng. Bull. **25**(1), 4–9 (2002)
3. Iannacone, M., Bohn, S., Nakamura, G., Gerth, J., Huffer, K., Bridges, R., Ferragut, E., Goodall, J.: Developing an ontology for cyber security knowledge graphs. In: Proceedings of the 10th Annual Cyber and Information Security Research Conference. pp. 1–4 (2015)

4. McGuinness, D.L., Van Harmelen, F., et al.: Owl web ontology language overview. W3C recommendation **10**(10), 2004 (2004)
5. Obrst, L., Chase, P., Markeloff, R.: Developing an ontology of the cyber security domain. In: STIDS. pp. 49–56 (2012)
6. Oltramari, A., Cranor, L.F., Walls, R.J., McDaniel, P.D.: Building an ontology of cyber security. In: STIDS. pp. 54–61. Citeseer (2014)
7. Scarpato, N., Cilia, N.D., Romano, M.: Reachability matrix ontology: A cybersecurity ontology. Applied Artificial Intelligence **33**(7), 643–655 (2019)
8. Sirin, E., Parsia, B.: Sparql-dl: Sparql query for owl-dl. In: OWLED. vol. 258 (2007)
9. Sokolov, I., Kupriyanovsky, V., Namiot, D., Sukhomlin, V., Pokusaev, O., Lavrov, A., Volokitin, Y.: Modern eu research projects and the digital security ontology of europe. International Journal of Open Information Technologies **6**(4), 72–79 (2018)
10. Syed, Z., Padia, A., Finin, T., Mathews, L., Joshi, A.: Uco: A unified cybersecurity ontology. In: Workshops at the Thirtieth AAAI Conference on Artificial Intelligence (2016)
11. Undercoffer, J., Joshi, A., Pinkston, J.: Modeling computer attacks: An ontology for intrusion detection. In: International Workshop on Recent Advances in Intrusion Detection. pp. 113–135. Springer (2003)

Применение объектной модели с интегрированным языком запросов для интеграции данных (Расширенные тезисы)

Владимир Ключиков

Московский Государственный Университет им. М.В. Ломоносова, Москва, Россия
kluchvlad@gmail.com

Аннотация. Модели представления данных можно разделить на два основных класса: реляционные (SQL) и нереляционные (NoSQL) модели данных. Модели данных классов сильно различаются, например, реляционные модели применяются для структурированных данных, а модели NoSQL в основном применяются для полуструктурированных данных. В работе решается проблема интеграции данных. Для решения вопросов интеграции данных необходимо найти модель, которая может объединить реляционные модели и модели NoSQL. Кандидатом в такую унифицирующую модель данных является объектный язык C# и его расширение – интегрированный язык запросов LINQ. Целью данной работы является разработка подхода к реализации интеграции данных с использованием сочетания языков C# и LINQ в качестве канонической модели данных.

Ключевые слова: интеграция данных, каноническая модель данных, объектная модель данных, интегрированный язык запросов.

При решении задачи интеграции данных необходимо разрешить проблемы неоднородностей на различных уровнях. На верхнем уровне преодолевается модельная неоднородность данных: данные можно разделить на реляционные и нереляционные (NoSQL). NoSQL модели подразделяются на различные категории: колоночные хранилища, документные хранилища, хранилища «ключ-значение», графовые СУБД и т.д[1].

Далее преодолевается неоднородность на уровне схем данных. Данные из различных источников могут быть представлены в различных схемах. При выполнении интеграции необходимо привести схемы источников к единой (целевой) схеме [2].

Для интеграции данных требуются следующие концептуальные спецификации: глобальная схема и отображения, которые связывают глобальную схему со схемами источников данных (локальных источников). Глобальная схема – это схема, которая является унифицированным представлением схем локальных источников.

[1] https://hostingdata.co.uk/nosql-database/

Существует несколько видов определения концептуальных отображений между глобальной схемой и локальными источниками. Среди них можно выделить следующие: Global-as-View (GAV) [5], Local-as-View (LAV) [5] и Global-Local-as-View (GLAV) [1, 3].

В интеграции данных каноническая модель играет роль унифицирующей (объединяющей) модели, в которой модели локальных источников могут быть представлены с сохранением семантики [7]. В качестве канонической модели данных в данной статье рассматривается сочетание объектного языка C# и интегрированного языка запросов LINQ[2]. В [4] со ссылкой на статью Мейера [6] приводятся основания, на основе которых предполагается, что сочетание C# и LINQ можно применять в качестве канонической модели для интеграции данных.

Цель настоящей работы заключается в разработке подхода к реализации системы интеграции данных, в котором сочетание объектного языка C# и интегрированного языка запросов LINQ используется в качестве канонической модели данных. Данная работа опирается на некоторые идеи и предположения, описанные в [4]. Подход заключается в разработке прототипа системы интеграции данных, в которой в качестве канонической модели выступает сочетание языков C# и LINQ.

Разработанный прототип состоит из набора связанных классов и интерфейсов, которые условно можно разделить на следующие группы: независимый интерфейс; классы, зависимые от предметной области; классы и интерфейсы, зависимые от видов источников; классы, зависимые от конкретных.

Вначале пользовательский запрос задается в «Классе пользователя» и посредством функции отправляется в «Класс регистрации источников, исполнения запроса и объединения результатов». Далее в «Классе регистрации источников, исполнения запроса и объединения результатов» выполняется регистрация источников с использованием конструкторов, определенных в каждом «Классе подключения к источнику». Затем пользовательский запрос вместе с параметрами соединения с каждым источником поочередно отправляется на переписывание в «Класс переписывания и исполнения запроса для источника». В методе, определенном в «Классе переписывания и исполнения запроса для источника» осуществляется определение взглядов (представлений), переписывание пользовательского запроса и исполнение запроса. Во время выполнения запроса осуществляется загрузка данных в прототип, и переписанный запрос исполняется локально над извлеченными из источника данными. После получения результатов исполнения запроса для каждого источника возвращаются в «Класс регистрации источников, исполнения запроса и объединения результатов» для объединения полученных результатов. Потом объединенные результаты из «Класса регистрации источников, исполнения запроса и объединения результатов» возвращаются в «Класс пользователя».

[2] https://docs.microsoft.com/ru-ru/dotneet/csharp/programming-guide/concepts/linq/

Основными чертами подхода являются динамическая материализация данных (данные загружаются из источников в память при исполнении запроса) и переписывание запросов с использованием техники GAV.

Разработанный прототип системы интеграции данных используется для интеграции источников, представленных в моделях, описанных в [4]. В прототипе интегрируются следующие источники, которые построены на основе выбранных моделей [4]: в качестве реляционной СУБД выбран SQLite[3], в качестве графового СУБД – Neo4j[4], в качестве документной – MongoDB[5], модели «ключ-значение» – Redis[6], колоночной модели – ClickHouse[7].

В качестве предметной области, на основе которой осуществлялось практическое применение прототипа, были выбраны статистические урбанистические данные [4]. Все источники хранятся в различных репозиториях: на локальном компьютере (данные по Барселоне), в песочнице на онлайн-сервере (Мадрид, Бильбао) или на виртуальной машине с ОС Linux (Севилья и Малага).

Рассматриваются следующие группы классов и интерфейсов: компоненты предметного посредника и компоненты, обеспечивающие связь предметного посредника с источниками.

В предметный посредник входит библиотека классов с описанием глобальной схемы, интерфейс исполнения запроса и реализующий данный интерфейс класс регистрации источников, исполнения запроса и объединения результатов, интерфейсы переписывания и исполнения запросов для каждого источника и реализующие данные интерфейсы классы переписывания и исполнения запросов для каждого источника.

Каждый из классов переписывания и исполнения запроса для каждого источника реализует интерфейс. В каждом из классов переписывания и исполнения содержится метод, принимающий на вход параметры подключения к источнику и непосредственно сам запрос. На основе параметров подключения осуществляется соединение с источником.

Во время выполнения команды исполнения запроса данные из источника материализуются, а переписанный запрос исполняется над загруженными из источника данными. Полученный результат возвращается в формате *List<object>*.

При построении взглядов возможно появление структурных конфликтов (конфликты имен, различные типы атрибутов …) и конфликтов значений (представление значений в различных единицах измерения).

Связь с источниками осуществляется путем определения классов подключения к источникам, а для связи прототипа с сущностями источника используются классы, в которых идет описание схемы локальных источников и их сущностей. В качестве примера в статье приведен запрос и результат исполнения запроса.

В описанном подходе максимальное использование возможностей C# и подключаемых библиотек приводит к ограничениям: Для каждого взгляда исполь-

[3] https://www.sqlite.org/index.html
[4] https://neo4j.com/
[5] https://www.mongodb.com/
[6] https://redis.io/
[7] https://clickhouse.tech/

зуются данные только из одного источника; Каждый источник должен покрывать все сущности глобальной схемы; При объединении результатов запросов к отдельным источникам не выполняется дедупликация данных; Подход не реализует в явном виде материализованную или виртуальную интеграцию, а сочетает некоторые их черты. Построение взглядов осуществляется вручную, а не автоматизированно.

Литература

1. Briukhov, D., Kalinichenko, L., Martynov, D.: Source Registration and Query Rewriting Applying LAV/GLAV Techniques in a Typed Subject Mediator. In: Proceedings of the 9th Russian Conference on Digital Libraries, RCDL'2007, pp. 253–262. Pereslavl, Russia (2007).
2. Briukhov D., Stupnikov S., Kalinichenko L., Vovchenko A.: Information Extraction from Multistructured Data and its Transformation into a Target Schema. In: Kalinichenko L., Starkov, S.: Selected Papers of the XVII International Conference on Data Analytics and Management in Data Intensive Domains (DAMDID/RCDL 2015), pp. 81–90 (2015).
3. Friedman, M., Levy, A., Millstein, T.D.: Navigational Plans for Data Integration. In: Proceedings of the National Conference on Artificial Intelligence (AAAI), pp. 67-73. AAAI Press/The MIT Press (1999).
4. Klyuchikov V.: Language Integrated Query as a Canonical Data Model for Virtual Data Integration. In: Elizarov A., Novikov B., Stupnikov S. (eds.): Proceedings of the XXI International Conference "Data Analytics and Management in Data Intensive Domains" (DAMDID/RCDL'2019), pp. 381–394, Kazan, Russia (2019).
5. Lenzerini, M.: Data Integration: A Theoretical Perspective. In: Proceedings of the 21st ACM SIGACT-SIGMOD-SIGART Symposium on Principles of Database Systems, pp. 233-246. Madison, Wisconsin, USA (2002).
6. Meijer, E., Bierman, G.: A co-Relational Model of Data for Large Shared Data Banks. Microsoft Research. ACMqueue 3(9), 1–19 (2011).
7. Stupnikov, S., Kalinichenko, L.: Extensible Unifying Data Model Design for Data Integration in FAIR Data Infrastructures. In: Manolopoulos Y., Stupnikov S. (eds): Data Analytics and Management in Data Intensive Domains (DAMDID/RCDL 2018). Communications in Computer and Information Science, vol 1003, pp. 17–36. Springer, Cham (2019).

Data Augmentation for Domain-Adversarial Training in EEG-based Emotion Recognition (Extended Abstract)

Ekaterina Lebedeva

Moscow State University, Moscow, Russia,
kate.1ebedeva@yandex.com

Emotion Recognition is an important and challenging task of modern affective computing systems. The results of EEG processing can be used in the creation of brain-computer interfaces (BCIs) and in neurophysiology studies. Solving this problem may also contribute the development of neuromarketing to determine consumer preferences [1]. And it has application in workload estimation and driving fatigue detection.

Neuronal action potentials measured by the Electroencephalography (EEG) provide an important data source with a high temporal resolution and direct relevance to a human brain activity. EEG is a multichannel continuous signal recorded with electrodes that measures differences between electric potentials. EEG-signal processing has a number of peculiarities. EEG recording is always contaminated with artifacts, such as EOG(ocular), ECG(cardiac), EMG(muscle), and noise. Different processes are reflected in different frequency bands of the electrical activity of the brain. As possible variants of the experimental protocol the following systems for recording EEG signals are used: Resting states with eyes open (REO) or with eyes closed (REC), Event-related potentials (ERPs), Task-related, Somnography.

Emotion is hard to measure because it is a subjective feeling. Emotions can be evaluated in terms of "positive", "negative" or "like", "dislike" [1]. Or there can be distinguished a set of basic emotions. Researchers often use a two- or three-dimensional space to model emotions [2, 3], where different emotion points can be plotted on a 2D plane consisting of a Valence axis and Arousal axis or on a 3D area with addition Dominance axis.

Machine-learning approach is one of the classic ways for solving the problem of emotion recognition. In this method it is necessary to extract reliable informative features closely related to the emotional state of the subject. The signal is divided into components by Independent Component Analysis (ICA) [5] to separate artifacts. The main method used for extracting spectral features is the Fourier transform [11]. In machine-learning approach, discriminant analysis [13] or Bayesian analysis [12] can be used for classification.

The more modern way of solving the problem is deep learning methods usage. This approach is often used as conjunction machine learning feature extraction with neural network classification.

Neural networks require a big amount of training data. EEG-based evaluation of the emotional state is complicated due to the lack of labeled training data

and to a strong presence of subject- and session-dependencies. Several datasets could be combined to increase the amount of training data. But it is difficult to conduct training on data from several sources because of different experiment conditions. Various adaptation techniques can be applied to train a model that would be robust to a domain mismatch in EEG data but the amount of available training data is still insufficient [15].

There are several datasets for EEG-based emotion recognition task. Every corpus was collected according to unique protocol. Available datasets for solving the problem are described in the paper. Such as DEAP dataset (A Database for Emotion Analysis Using Physiological Signals) [7], eNTERFACE-2006 project [8], SEED (SJTU Emotion EEG Dataset) [14] and SEED-IV [9] datasets, The Neuromarketing dataset [1] and Imagined Emotions [10]. There is described the number of participants of each experiment, stimuli, emotion measurement and data format.

The article provides an overview of related works on each stage of EEG data processing. Electroencephalogram data has a low signal to noise ratio. Therefore the extensive filtering and artifact removal procedures must be included as a necessary part of the analysis pipeline. After the cleaning step, the multichannel signal can be decomposed into quasi-independent components by the Independent Component Analysis (ICA) or with more recent autoencoder-based approaches.

During the feature extraction step, EEG signal is divided into short time frames. Feature extraction is performed independently for each frequency band of signal frames. Such metrics and statistics can be utilized as informative features: max, min, average amplitude and Power spectral density (PSD). Following cross-channel features can be calculated: Root Mean Square, Pearson Correlation Coefficient between 2 channels, Magnitude Squared Coherence Estimate.

Emotion recognition problem in feature space can be approached with one of the machine learning methods for classification. The article discusses works that suggest using such classification methods as Naive Bayes model [12], K Nearest Neighbours classifier and Linear Discriminant Analysis [13].

In the studied area, deep neural network-based feature extraction and emotion recognition began to be intensively applied. The examples of using the following neural network architectures are given in the article: Deep Belief Network (DBN) [14] trained with differential entropy features, Stacked autoencoder as a substitute for Independent Component Analysis, LSTM-RNN for classification based on Frequency Band Power Features extracted from the SAE output [6].

It is important to apply a domain adaptation technique to a model that would compensate the subject variability or heterogeneity in various technical specifications. The paper [15] compares different domain adaptation techniques on two datasets: DEAP and SEED. Transfer Component Analysis (TCA) and Maximum Independence Domain Adaptation (MIDA) performed the best results for subject within-dataset domain adaptation. In [16] another approach to a domain adaptation was considered based on neural networks which are optimized by minimizing the classification error on the source, while making the source and

the target similar in their latent representations. Where source and target are two inputs of data from different subjects. This method was compared with multiple domain adaptation algorithms on benchmark SEED and DEAP.

In this work we propose an approach based on the domain adversarial training and combining available training corpus with much larger unlabeled dataset in a semi-supervised training framework. There are presented two architectures of the approach: with one and two inputs. These architectures differ in that in the first case, the classification of domains occurs independently for input samples, and in the second, pairwise comparisons are performed. In the future work, a comparison of these two approaches will be carried out and the best approach will be defined.

In order to improve the performance of the model, a large number of EEG datasets without affective labels can be utilized emotion recognition task. DEAP dataset includes data of only 32 subjects, as well as other datasets for EEG-based emotion recognition also contain a limited variability of subjects. It is more efficient to train neural networks on such data volume as the Temple University Hospital (TUH) EEG data corpus [17] with more than 10000 participants, therefore, as the solution it is proposed to use unlabeled data. Thus, the neural network will be trained on a larger set of subjects, and therefore, will provide a better generalized model for new subjects.

It is possible to use EEG datasets without emotional labels if they contain video recordings of the experiment. Data can be marked by emotions detected from the video. Unfortunately, the EEG datasets with the recordings of such modalities are rare, so this approach probably will not be allow to significantly expand the training data.

The effect of emotion recognition performance degradation caused by the subject- and session-dependencies was measured on DEAP dataset proving the need to develop approaches that would utilize larger datasets in order to obtain a better generalized model.

This paper discribes the EEG-based emotion recognition task and its existing solution methods. There was formulated the problem of domain mismatch and insufficient data amount for training neural networks. As a solution, there was proposed the application of existing domain adaptation techniques with data augmentation due to datasets without emotional labels.

References

1. Yadava, M., Kumar, P., Saini, R. et al. Analysis of EEG signals and its application to neuromarketing. Multimed Tools Appl 76, 1908719111 (2017). https://doi.org/10.1007/s11042-017-4580-6
2. Lang, P. J. (1995). The emotion probe. Studies of motivation and attention. Am.Psychol. 50, 372385.
3. Scherer, K. R. (2005). What Are Emotions? And How Can They Be Measured? Social Science Information, 44, 695-729. https://doi.org/10.1177/0539018405058216
4. Robert Jenke, Angelika Peer, and Martin Buss. Feature extraction and selection for emotion recognition from EEG. IEEE Transactions on Affective Computing, 5(3):327339, jul 2014.

5. A. Hyvarinen and E. Oja, Independent Component Analysis: A Tutorial, LCIS, Helsinki University of Technology, Finland. April 199
6. Xing X, Li Z, Xu T, Shu L, Hu B and Xu X (2019) SAE+LSTM: A New Framework for Emotion Recognition From Multi-Channel EEG. Front. Neurorobot. 13:37. https://doi.org/10.3389/fnbot.2019.00037
7. Koelstra, S., Muhl, C., Soleymani, M., Lee, J. S., Yazdani, A., Ebrahimi, T., ... & Patras, I. (2011). Deap: A database for emotion analysis; using physiological signals. IEEE transactions on affective computing, 3(1), 18-31.
8. Savran, A., Ciftci, K., Chanel, G., Mota, J., Hong Viet, L., Sankur, B., ... & Rombaut, M. (2006). Emotion detection in the loop from brain signals and facial images. In Proceedings of the eNTERFACE 2006 Workshop.
9. Wei-Long Zheng, Wei Liu, Yifei Lu, Bao-Liang Lu, and Andrzej Cichocki, EmotionMeter: A Multimodal Framework for Recognizing Human Emotions. IEEE Transactions on Cybernetics, Volume: 49, Issue: 3, March 2019, Pages: 1110-1122, https://doi.org/10.1109/TCYB.2018.2797176
10. Onton J and Makeig S (2009). High-frequency broadband modulations of electroencephalographic spectra. Front. Hum. Neurosci. 3,:61. https://doi.org/10.3389/neuro.09.061.2009
11. Grass, A. M. & Gibbs, F. A. (1938). A FOURIER TRANSFORM OF THE ELECTROENCEPHALOGRAM. Journal of Neurophysiology, 1(6), 521–526. https://doi.org/10.1152/jn.1938.1.6.521
12. N. Sebe, M. S. Lew, I. Cohen, A. Garg, and T. S. Huang, Emotion recognition using a cauchy naive bayes classifier, in Proc. ICPR, vol. 1, 2002, pp. 1720.
13. Murugappan, M., Ramachandran, N. & Sazali, Y. (2010). Classification of human emotion from EEG using discrete wavelet transform. Journal of biomedical science and engineering, 3(04), 390.
14. Wei-Long Zheng & Bao-Liang Lu. (2015). Investigating Critical Frequency Bands and Channels for EEG-Based Emotion Recognition with Deep Neural Networks. IEEE Transactions on Autonomous Mental Development, 7(3), 162175. https://doi.org/10.1109/tamd.2015.2431497
15. Lan, Z., Sourina, O., Wang, L., Scherer, R. & Mller-Putz, G. R. (2018). Domain adaptation techniques for EEG-based emotion recognition: a comparative study on two public datasets. IEEE Transactions on Cognitive and Developmental Systems, 11(1), 85-94.
16. Li, J., Qiu, S., Du, C., Wang, Y. & He, H. (2019). Domain Adaptation for EEG Emotion Recognition Based on Latent Representation Similarity. IEEE Transactions on Cognitive and Developmental Systems, 11. https://doi.org/10.1109/TCDS.2019.2949306
17. Obeid I and Picone J (2016) The Temple University Hospital EEG Data Corpus. Front. Neurosci. 10:196. https://doi.org/10.3389/fnins.2016.00196

One- and Unidirectional Two-dimensional Signal Imitation in Complex Basis (Extended Abstract)

Ivan Deykin[0000-0002-8151-6337]

Bauman Moscow State Technical University, Moscow 105005, Russia
deykinii@student.bmstu.ru

Abstract. Signal imitation is widely used today since it helps to bring the experiment to the virtual domain thus eliminating risks of damaging real equipment. At the same time all signals used in the physical world are limited by the finite band of frequencies rendering bandpass signal studies especially important. The method for imitating bandpass signals in complex basis is favorable in the case of a bandpass signal as it uses resources effectively and provides the desired accuracy.

The author has implemented the method in the form of the PC application generating signals according to the characteristics set by the user. These characteristics are: borders defining the signal's frequency band, the time period, the number of steps for discretization, the spectral density form. The PC application uses the characteristics to generate the signal and its experimental autocorrelation. The application calculates theoretic and algorithmic autocorrelations in order to evaluate the quality of the imitation by computing the error function. The application visualizes all the resulting information via the simple interface.

The application was used to generate two-dimensional signals to highlight the present limitations and to sketch the direction for the future. The application is later to be adapted completely to imitating multidimensional signals.

This work is financially supported by the Russian Federation Ministry of Science and Higher Education in the framework of the Research Project titled "Component's digital transformation methods' fundamental research for micro- and nanosystems" (Project #0705-2020-0041).

Keywords: digital signal processing, DSP, Fourier functions, two-dimensional signals, broadband signal, signal imitation, random signal generation

1 Introduction

The word "signal" is widely used today. Temporal changes of some physical value can be represented as time series or as signals. The term "signal processing" is applicable to any processes that change in time [1, 2] including the very large time series data [3]. The fundament of such analysis is derived from the theory of digital signal processing [1].

Increasing volumes of data mean that the dimensions of signals grow too [4]. Multidimensional signals are used more and more [5, 6]. The complex basis was found useful for imitating one-dimensional bandpass signals [7, 8]. The program that uses the complex basis was designed and tested on one-dimensional (section 2) and two-dimensional unidirectional signals (section 3). It is planned to upgrade the designed program for imitating multidimensional signals with varying numbers of dimensions.

2 One-dimensional signal imitation in complex basis

Bandpass signal's spectrum is constrained within two border frequencies [8]. The goal of the imitation is to generate the signal fitting such a spectrum[9]. User inputs the form of the spectrum, its limiting frequencies ω_L and ω_R, the period T and the number of discretization intervals N [10]. "L" stands for "Left" and "R" stands for "Right". Three the signals and also three experimental autocorrelations generated by the program are presented on the figure 1.

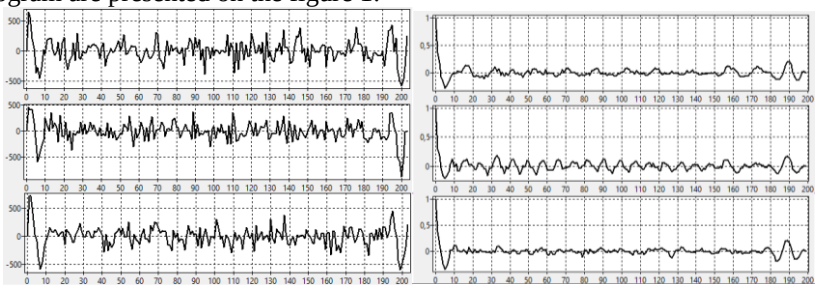


Fig. 1. Three random signals (on the left) and three resulting autocorrelations (on the right) based on the same spectral density

3 Two-dimensional signal imitation in complex basis

The structure of multidimensional signals presents the certain level of difficulty [4]. Before advancing into two-dimensional domain it was decided to study the "quasi-two-dimensional" signals that are obtained by stacking together one-dimensional broadband signals generated earlier. Spectral density of such a "quasi-two-dimensional" signal is presented on the figure 2.

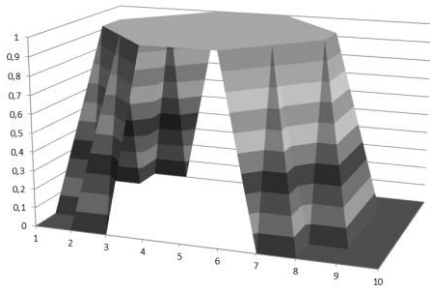


Fig. 2. Two-dimensional spectral density

Then the one-dimensional signals comprising the two-dimensional one can be generated separately and stacking them together side by side provides us with a two-dimensional signal (figure 3).

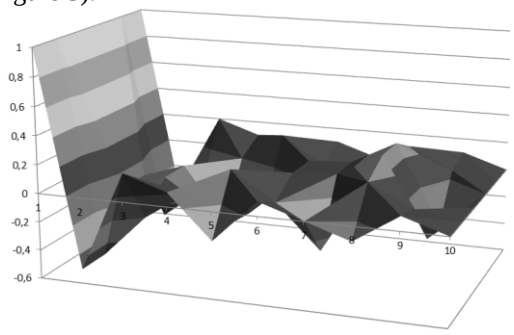


Fig. 3. Two-dimensional signal imitated

Signals generated look like unidirectional waterfall plots which are encountered in many different fields [11, 12, 13], so the need for generating such signals is present.

4 Conclusion

This paper is a part of a new development for high-dimensional signal simulation that is presented in the conference by the paper where the author was involved too. The method of imitation in complex basis realized earlier and used here to imitate two-dimensional signals is preferential for its computational complexity and scalability.

The software solution was implemented in the Lazarus IDE to meet all the criteria and to ensure project's applicability within the educational process. The visualization in the section 3 was obtained using MS Excel 2010.

The first test of the one-dimensional algorithm being used to imitate two-dimensional signals highlighted the direction for future development: the algorithm should be adopted for signals of different dimensions; the visualization facilities should be expanded. The method as it is can be used for modeling the unidirectional two-dimensional data in the form of a waterfall plot.

Acknowledgements

This work was supervised by professors Syuzev V. V. and Smirnova E. V. of the Bauman Moscow State Technical University. The project was executed with the financial support of the Russian Federation Ministry of Science and Higher Education in the framework of the Research Project titled "Component's digital transformation methods' fundamental research for micro- and nanosystems" (Project #0705-2020-0041). Special gratitude goes to the organizers of DAMDID conference for providing a medium suitable for exchanging ideas and results and advancing the quality of scientific work.

References

1. Richard G. Lyons. Understanding Digital Signal Processing (3rd Edition). Prentice Hall. 2010.
2. Ceri S. et al. (2018) Overview of GeCo: A Project for Exploring and Integrating Signals from the Genome. In: Kalinichenko L., Manolopoulos Y., Malkov O., Skvortsov N., Stupnikov S., Sukhomlin V. (eds) Data Analytics and Management in Data Intensive Domains. DAMDID/RCDL 2017. Communications in Computer and Information Science, vol 822. Springer, Cham.
3. Kraeva Y., Zymbler M. (2019) Scalable Algorithm for Subsequence Similarity Search in Very Large Time Series Data on Cluster of Phi KNL. In: Manolopoulos Y., Stupnikov S. (eds) Data Analytics and Management in Data Intensive Domains. DAMDID/RCDL 2018. Communications in Computer and Information Science, vol 1003. Springer, Cham.
4. Bourennane, Salah & Marot, Julien & Fossati, Caroline & Bouridane, Ahmed & Spinnler, Klaus. (2014). Multidimensional Signal Processing and Applications. TheScientificWorldJournal. 2014. 365126. 10.1155/2014/365126.
5. Andrew S. Glassner. Principles of Digital Image Synthesis. Morgan Kaufmann Publishers, inc. San Francisco, California. USA. 1995. ISBN 1-55860-276-3.
6. Kosiński, K., Stanek, J., Górka, M.J. et al. Reconstruction of non-uniformly sampled five-dimensional NMR spectra by signal separation algorithm. J Biomol NMR 68, 129–138 (2017). https://doi.org/10.1007/s10858-017-0095-8
7. Deykin I., Syuzev V., Gurenko V., Smirnova E., Lubavsky A. (2019) Random Bandpass Signals Simulation with Complex Basis Algorithm / Science Problems Journal. 2019. № 11 (144).
8. Syuzev V.V., Smirnova E.V., Gurenko V.V. Speed Algorithms for Signal's Simulation // Science Problems. 2018. №11 (131). URL: https://cyberleninka.ru/article/n/bystrye-algoritmy-modelirovaniya-signalov (Date of retrieve: 04.05.2020) – in russian.
9. Syuzev V.V., Gurenko V.V., Smirnova E.V. Signal Simulation Spectra Algorithms as Learning and Methodical Tools of Engineers' Preparation // Machinery and Computer Technologies. 2016. №7. URL: https://cyberleninka.ru/article/n/spektralnye-algoritmy-imitatsii-signalov-kak-uchebno-metodicheskiy-instrument-podgotovki-inzhenerov (Date of retrieve: 04.05.2020).
10. Kotelnikov, V. A. On the carrying capacity of the ether and wire in telecommunications. Material for the First All-Union Conference on Questions of Communication, Izd. Red. Upr. Svyazi RKKA. 1933. p. 769 – 770. – in russian.
11. Johannesen, L. & Grove, U.S.L. & Sørensen, J.S. & Schmidt, M. & Graff, Claus & Couderc, Jean-Philippe. Analysis of T-wave amplitude adaptation to heart rate using RR-binning of long-term ECG recordings. Computing in Cardiology. 37. 369 - 372. 2010.
12. Marinelli, Eugene & Hyde, Truell & Matthews, Lorin. The Effects of Particle Size and Polydispersion on Complex Plasma Crystal Structure and Phase Transition. 2020.
13. Xu, Xiaochi & Vallabh, Chaitanya Krishna & Cleland, Zachary & Cetinkaya, Cetin. Phononic Artifacts for Real-time In-Situ Quality Monitoring of 3D Printed Objects at Fiber/Bond-scale. Journal of Manufacturing Science and Engineering. 10.1115/1.4036908. 2017.

Analysis of Gaze Trajectories in Natural Reading with Hidden Markov Models (Extended Abstract)

Maksim Volkovich

Lomonosov Moscow State University

1 Introduction

Natural reading is a complex task that includes eye movement processes, lexico-semantic processing dependent on reader attention and visual features of the text being read. The process of reading consists of long relatively rare movements ("saccades") between areas of high attention where eyes are fixated on a word for some time depending on the skill of reader. Eye movements during reading are under the direct control of linguistic processing [1]. There are three properties of a word that influence its ease of processing: word's frequency, length and predictability in context, the so-called Big Three [2]. Medical conditions such as schizophrenia and dyslexia also influence eye-movements during reading. [3–5]. These words properties and reader's conditions affect time of word's processing, length of saccade and probability of word's skipping.

Eye-tracking signals and features extracted from eye-tracker are successfully used in a large number of different tasks such as reader's attention classification [6], reader's language proficiency determining [7], part-of-speech tagging [8], named entity recognition [9]. Different variations of Hidden Markov Models were used before in order to determine the part of speech, the correct coordinate at which the eye is directed etc. In this work we propose to model eye-tracking trajectories using Hidden Markov Models: associate a set of hidden states with individual read words and fit the matrix of transition probabilities between them. We assume that this matrix can be used as a set of features for a models solving various cognitive state recognition tasks.

For future studies and experiments two publicly available existing eye-tracking datasets were chosen: Zurich Cognitive Language Processing Corpus [10] and Ghent Eye-Tracking Corpus [11]. Chosen datasets include such statistics of gaze trajectories during natural reading as time after first fixation on sentence for every single fixation, number of fixations for each word and sentence and mean pupil size, gaze duration (GD) during first word reading, sum of reading time of word (total reading time or TRT), first fixation duration (FFD), the duration of first and single fixation on a word (single fixation duration or SFD) and go-past time (GPD) which is the sum of all fixations preceding saccade to the right.

2 Proposed Approach

A first-order Hidden Markov Model is a probabilistic model which is based on Markov Chain model. An HMM is defined by the sequence of observed values, a sequence of hidden states that correspond to observed values, a transition probability matrix, observation likelihoods or emission probabilities, expressing probabilities that observed value would be generated from a hidden state and an initial probability distribution. A first-order HMM instantiates 2 assumptions. The first one is Markov assumption: the value of the hidden state depends only on the state at the previous moment. Second, the value of the observed value depends only on the current hidden state. It is known as Output Independence assumption. Having a sequences of observed values (in our case, coordinates taken from an eye-tracker) we need to determine from which hidden states (read words) these coordinates were led from. It is needed to solve HMM learning problem: learn the transition probability matrix and observation likelihoods from given observation sequences and set of possible hidden states where each hidden state would represent read word. For an HMM model parameters Baum-Welch algorithm (the special case of estimation Expectation-Maximization or EM algorithm) can be applied.

Our primary goal is to propose an approach for extracting subject-dependent features from eye-tracking data that would correspond to cognitive states and a group of text-dependent features. Examples of cognitive states that can be extracted are a level of fatigue, stress, focus, a level of text understanding, emotional state. One of the features of the text can be "difficult" words that take a longer reading time, require higher number of fixations and longer fixations that can be explained by the requirements for the level of proficiency in the language or subject area. Or we can try to discern the influence of words which are unpredictable in context on eyes movements patterns. From practical point of view, one of two following situations is considered for further analysis: either a same text is read by a certain number of different subjects, or a single subject reads a set of texts of sufficient size. Thus, two different models are proposed for these scenarios, text-dependent and subject-dependent.

In a "same text, different readers" scenario, text-dependent features remain the same for every session, but reader-dependent features should be different for different readers. The main objective in the analysis of this scenario is to determine a vector of text-dependent parameters given a set of observation sequences related to one text fragment read by a certain number of subjects. It is assumed that a set of observations related to a large enough set of subjects can be used to estimate a subject-independent HMM parameters which can be analysed for the purpose of extracting text-related features while subject-dependent features would be suppressed. Parameters of the model trained from samples taken from different subjects reading the same text may potentially represent such text characteristics as sentence structure, word frequency, general text complexity, etc.

In a scenario when a single subjects reads a certain number of text fragments it may be assumed that during one session cognitive states of a subject should

not change significantly over time and therefore model parameters should be similar for every single page. The objective of the analysis in this scenario is to estimate these subject-dependent parameters. Since the data is presented as a set of eye-tracker trajectories collected from different text fragment, each single trajectory is assumed to be sampled from a corresponding text-dependent HMM. It is assumed that parameters of each HMM can be presented as a function of higher-level parameters that refer to current cognitive state of a reader. For example, a level of fatigue can be modeled as the average duration of a fixation on a word. In order to train a subject-dependent HMM on these data, a parametric family of HMM is proposed.

3 Experiments

A set of experiments was executed to prove a concept that eye-tracking trajectories can be generated using HMM transition probability matrix and then HMM can be fitted the model parameters can be fitted close to the parameters of the original model. For a given set of sentences we can obtain coordinates of exact positions of words on a page. The distance was measured in relation to the size of a printed letter. A Hidden Markov Model was initialized in the following way according to our vision of what they should look like. The number of hidden states was set according to the number of displayed words. A transition probability matrix was generated as diagonally dominant since the number of saccades and therefore hidden states transitions has to be much smaller then number of eye movements inside areas of gaze fixations on a single word when the hidden state does not change. Initial probability distribution was chosen to be geometric distribution in order to simulate a task in which the subject must read the text from the beginning. A set of gaussian 2-d distributions with means located in word centers was chosen as a set of emission probabilities.

Mean Squared Error over the difference of original and a trained transition probability matrices was chosen as a metric. For a training over dataset consists of sequences with 300 observed values the chosen metric reaches values close to optimal at observations number between 10 and 15 sequences. With a further increase in observations numbers MSE over full transition probability matrices and MSE over diagonals do not decrease significantly. So for a suboptimal quality it could be enough to collect data from dozen of subjects.

4 Conclusion and Future Work

In this work various approaches to the analysis of eye movement in different text-dependent and subject-dependent scenarios have been considered. A new approach of gaze trajectories modeling based on Hidden Markov Models is proposed for both scenarios. An experiment conducted as part of a preliminary study helped us determine the approximate size of the sample we would need for a further work. It is planned to apply a proposed approach on real data taken from ZuCo and GeCo datasets in one of the following tasks.

1. Binary classification task: was the answer given by subject in a sentiment task from ZuCo dataset was correct or not.
2. If information about the answer time is available it may be possible to evaluate this parameter using a model, because it is supposed that time required for answer is dependent on how thoroughly the text was read.
3. The fatigue classification task. It is assumed that at the end of the long reading session a fatigue level is higher then at the beginning of the session. Data from GeCo corpus would be useful.

It is also planned to collect a new dataset containing samples of eye movement recordings taken while reading text fragments in Russian. Several dozen subjects will read several texts at different levels of fatigue. Fatigue level will be measured in two ways: by interviewing subjects and using a binary classifier trained to recognize the beginning and end of the session. Using the second method, we will assume that fatigue level rises at the end of the session.

References

1. Dambacher, M., Slattery, T. J., Yang, J., Kliegl, R., & Rayner, K. (2013). Evidence for direct control of eye movements during reading. Journal of Experimental Psychology: Human Perception and Performance, 39(5), 1468.
2. Clifton Jr, C., Ferreira, F., Henderson, J. M., Inhoff, A. W., Liversedge, S. P., Reichle, E. D., & Schotter, E. R. (2016). Eye movements in reading and information processing: Keith Rayners 40 year legacy. Journal of Memory and Language, 86, 1-19.
3. Whitford, Veronica, et al. "Reading impairments in schizophrenia relate to individual differences in phonological processing and oculomotor control: Evidence from a gaze-contingent moving window paradigm." Journal of Experimental Psychology: General 142.1 (2013): 57. APA
4. Olson, Richard K., Reinhold Kliegl, and Brian J. Davidson. "Dyslexic and normal readers' eye movements." Journal of Experimental Psychology: Human Perception and Performance 9.5 (1983): 816.
5. De Luca, Maria, et al. "Eye movement patterns in linguistic and non-linguistic tasks in developmental surface dyslexia." Neuropsychologia 37.12 (1999): 1407-1420.
6. Mozaffari, Seyyed Saleh, et al. "Reading Type Classification based on Generative Models and Bidirectional Long Short-Term Memory." (2018).
7. Kunze, Kai, et al. "Towards inferring language expertise using eye tracking." CHI'13 Extended Abstracts on Human Factors in Computing Systems. 2013. 217-222.
8. Barrett, Maria, et al. "Weakly supervised part-of-speech tagging using eye-tracking data." Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers). 2016.
9. Hollenstein, Nora, and Ce Zhang. "Entity Recognition at First Sight: Improving NER with Eye Movement Information." arXiv preprint arXiv:1902.10068 (2019).
10. Hollenstein, Nora, et al. "ZuCo, a simultaneous EEG and eye-tracking resource for natural sentence reading." Scientific data 5.1 (2018): 1-13.
11. Cop, Uschi, et al. "Presenting GECO: An eyetracking corpus of monolingual and bilingual sentence reading." Behavior research methods 49.2 (2017): 602-615.
12. hmmlearn library, https://hmmlearn.readthedocs.io

Применение методов машинного обучения для кросс-идентификации астрономических объектов (Расширенные тезисы)

Александра Кулишова[1]

[1] Московский государственный университет им. М. В. Ломоносова, Москва, Россия
sasha_kulishova@mail.ru

Одной из важных задач в астрономии является задача кросс-идентификации астрономических объектов, т.е. задача отождествления наблюдений с объектами, к которым они относятся [5]. Эти объекты в зависимости от системы могут принадлежать, как к одному каталогу, так и к разным. Кросс-идентификация необходима, но имеет ряд сложностей, связанных с различием методов получения информации для астрономических каталогов. Кроме того, эта информация также зависит от чувствительности инструментов наблюдения, которая дает различную астрометрическую и фотометрическую погрешность. Эти различия приводят к тому, что один и тот же физический источник в разных каталогах может иметь отличающиеся характеристики/

Одним из наиболее известных методов является Байесовский подход к кросс-идентификации астрономических объектов. Однако, с ростом популярности методов машинного обучения, появились работы, использующие их в данной задаче. Это обусловлено тем, что они могут работать быстрее и с большим объемом данных.

Цель данной работы - применение методов машинного обучения для кросс-идентификации астрономических объектов. В рамках статьи были выполнены следующие подзадачи:

1. Анализ существующих подходов кросс-идентификации астрономических объектов на основе методов машинного обучения.
2. Выбор подходящих методов машинного обучения на основе результатов анализа.
3. Реализация выбранных методов: выбор данных, проведение экспериментов, выявление особенностей реализации.
4. Сравнительный анализ выбранных методов на наборах астрономических данных Gaia и PLAsTiCC.

Чтобы проанализировать существующие подходы кросс-идентификации астрономических объектов с помощью методов на основе машинного обучения было рассмотрено несколько схожий работ. Так, в статье “Cross-matching Gaia objects” предлагается разделить задачу перекрестного сопоставления для анализа данных в Gaia на два этапа: классификации и кластеризации. Задача классификации возникает, когда нам заранее задан некоторый каталог или несколько каталогов источников, которые мы сопоставляем между собой либо к которым соотносим поступающие наблюдения. Этап кластеризации необходим при

условии, что у нас нет исходного каталога. Для задачи классификации автор использует алгоритм К-ближайших соседей, а для этапа кластеризации предлагает, как и в [1], [3], использовать на данном этапе агломеративную кластеризацию. Более подробно этап классификации рассматривается в статье [4]. Здесь авторы предлагают использовать нейронные сети и метод опорных векторов для перекрестного сопоставления объектов и доказывают, что эти методы могут показать хорошие результаты.

В рамках работы из каждого каталога было выбрано одинаковое количество наблюдений (10000), к которым независимо применялись алгоритмы классификации и кластеризации. Этот процесс состоял из нескольких этапов: предобработка данных и разделение их на обучающую и валидационную выборки, подбор параметров и обучение модели, получение показателей качества модели и сравнение результатов.

Проверка эффективности применения рассмотренных методов машинного обучения проводилась на второй версии данных наблюдений Gaia (data release 2) [6] и данных соревнований PLAsTiCC [7]. Они отличаются по способу сбора информации, признакам, описывающим каждое наблюдение, и количеству данных.

Так как используемые каталоги созданы для решения различных астрономических задач, во время предобработки был выполнен отбор необходимых для текущей задачи признаков. Кроме того, чтобы приблизить данные к реальным условиям, был дополнительно сгенерирован шум для некоторых важных признаков, которые в текущих каталогах были заданы константными значениями, например, координаты. После этапа предобработки данные были разделены на обучающую и валидационную выборки в соотношении 7:3, а также и 2:8 для дополнительной проверки.

Для выбора наилучшего метода использовалось несколько подходов. Для задачи классификации изначально проводилась перекрестная проверка, как и в статье [4]. Далее метода проверялись на обучающей и валидационных выборках - на первой модель обучалась, на второй считались выбранные показатели качества. Кластеризация проводилась на всех 10000 наблюдений сразу, а далее сравнивалась с набором истинных меток с помощью выбранных показателей качества.

В результате на данных PLAsTiCC лучше всего себя показали решающее дерево и многослойный персептрон - доля правильных ответов этих моделей при перекрестной проверке и при разделении данных в отношении 7:3 превысила 97%. Этот же показатель качества у оставшихся алгоритмов колебался около 92%. Чтобы проверить корректность полученных результатов, было решено применить модели к выборкам, сформированных в соотношении 2:8. На этом этапе были получены схожие результаты, то есть уменьшение в 3.5 раза обучающего набора не помешало алгоритмам качественно классифицировать данные (с точностью более 80%).

На данных Gaia модели показали результат хуже, чем на данных из PLAsTiCC. Это связано со значением величины потока и смежными с ним признаками. Лучше всего сработали решающее дерево и К-ближайших соседей с долей правильных ответов около 72% и 55%. Показатели метода опорных век-

торов: доля правильных ответов около 25%, F1-мера – 40%. F1-мера многослойного персептрона - 15%. Результаты многослойного персептрона и SVM являются недостаточным для поставленной задачи. Возможно, это связано с силой шумов при измерении потока, но решение данной проблемы будет обсуждаться в следующей работе.

Что касается кластеризации, то лучше всего себя показала агломеративная кластеризация. Уровни показателей индекса Рэнда и V-меры на данных PLAsTiCC связаны с низким уровнем однородности, т.е. почти все наблюдения одного и того же объекта попадают в один кластер, но в этот же кластер попадают наблюдения другого источника. Это связано с близким расположением этих объектов. На данных Gaia алгоритмы кластеризации повели себя аналогично методам классификации: большое количество наблюдений соотносились к одному кластеру. Но при дополнительной предобработке, результат становился аналогичным результату при работе с данными PLAsTiCC.

Полученные результаты показали, что методы машинного обучения могут быть применены к задаче кросс-идентификации объектов и даже дать хороший результат. Одним из таких алгоритмов является решающее дерево, которое не требует ни нормализации, ни масштабирования данных, а реализация его достаточно проста.

Однако структура данных и их качество все еще влияют на работу алгоритмов и требуют более высокого уровня этапа предобработки. В будущей работе необходимо будет решить проблемы, связанные с влиянием признака величины потока в данных Gaia, формированием обучающего и валидационного наборов. Также в последующей работе будет рассмотрено применение к данной задачи таких алгоритмов, как: XGBoost, LightGBM и RandomForest. А вместо MLPClassifier будет построена полноценная нейронная сеть с использованием библиотеки keras. Кроме того, будет проведена работа по исследованию применения метрик расстояния в задаче кластеризации с целью повышения качества работы рассматриваемых алгоритмов.

Литература

1. M. Clotet, J. Castaneda, J. Torra and other: Cross-matching algorithm for the intermediate data updating system in Gaia (2016).
2. L. Lindegren: Cross-matching Gaia objects (2015).
3. F. Torra, M. Clotet, J. J. Gonzalez-Vidal and other: Proper motion and other challenges in cross-matching Gaia observations (2018).
4. D. J. Rohde, M. J. Drinkwater: Applying machine learning to catalogue matching in astrophysics (2005).
5. Малков О.Ю., Длужневская О.Б., Кайгородов П.В., Ковалева Д.А., Скворцов Н.А.: Проблемы обозначения и кросс-идентификации кратных объектов в астрономии (2015).
6. Lindegren L.: Gaia Data Release 2: The astrometric solution (2018).
7. Tarek Allam Anita Bahmanyar, Rahul Biswas и другие: The Photometric LSST Astronomical Time-series Classification Challenge (PLAsTiCC): Data set (2018).

Machine Learning Models in Predicting Hepatitis Survival Using Clinical Data*
(Extended Abstract)

Kouame Amos Brou[0000−0003−1996−577x]

Peoples' Friendship University of Russia, Moscou 117198, Russia
broukouameamos9@gmail.com

Hepatitis is an inflammation of the liver. There are two main categories of hepatitis. Viral hepatitis, caused by a virus, and non-viral hepatitis, mainly caused by the ingestion of products toxic to the liver due to alcohol, toxic chemicals, etc... Non-viral hepatitis can also be the result of diseases affecting the liver, such as liver steatosis and autoimmune hepatitis. The most common symptoms are: loss of appetite, muscle and joint pain, weight loss, fatigue, insomnia, hypersomnia, nausea, vomiting, diarrhea, headaches, yellowing of the eyes etc....

Hepatitis kills about 1.34 million people a year, a number comparable to deaths from tuberculosis and AIDS. This mortality due to hepatitis is increasing while that due to tuberculosis or AIDS is decreasing, noted Dr. Gottfried Hirnschall, director of the hepatitis program at the WHO. The proposed model is a combination of rules and different machine learning techniques. Machine learning models can help physicians reduce the number of false decisions. In this article, we will discuss the construction of a predictive model based on machine learning that can predict the survival of a hepatitis patient from medical data.

The objective of this study is to propose a classification model based on rules and machine learning techniques for the prediction of patient survival. For this we used a dataset of 308 patients in which 278 patients (90.26%) and 48 patients (9.74%) were female and male respectively. The different models of machine learning namely: Random Forests (RF), K-Nearest Neighbors (KNN), Linear Support Vector Machine (LSVM), Logistic Regression (LR), Decision Tree (DT), Naive Bayes (NB) and AdaBoost (AB) as well as the 10-folds cross validation technique were used for this study. Some performance measures of machine learning such as Accuracy, precision, Sensitivity, Specificity and ROC curve were evaluated.

Out of 308 patients, 244 patients and 64 patients were alive and dead respectively. The standardized confusion matrix shows that the accuracy of KNN, LR, LSVM, RF, NB, DT, AB is 96%, 77%, 78%, 96%, 65%, 98.70%, 91%. In this study, the Decision Tree (DT) shows better results compared to the other models (KNN, LSVM, LR, NB, RF, AB). Accuracy, sensitivity and RF ROC curve are respectively 98.70%, 99%, 98%. However, the NB has low accuracy 65%, sensitivity 58% and ROC curve 70%..

* The publication has been prepared with the support of the \RUDN University Program 5-100" and Ivan Valentinovich Smirnov (ivs@isa.ru)

In this study on the prediction of survival in hepatitis patients, the Decision Tree proved to be the most efficient classification model. Therefore, this model can be recommended as a useful tool for the prediction of survival of hepatitis patients as well as for medical decision making. This method is quick to apply because it is resistant to training and the ability to manage data with and without pre-processing, for example, the data does not need to be resized, transformed or modified. It can also be used for feature selection (searching for effective risk factors) only. The analysis of this study showed that the absence of variables such as steroid, antiviral, liver-big and liver-firm do not significantly affect the prediction result. However, the increase in our dataset data could allow our model to have better performance than those presented above.

Keywords: Predictability of hepatitis · Classification · Machine learning models.

Алгоритм автоматической акцентуации с учетом орфоэпической нормы автора* (Расширенные тезисы)

А. В. Мосолова

Институт вычислительных технологий СО РАН
a.mosolova@g.nsu.ru

Во многих областях анализа текста для его корректного рассмотрения требуется наличие ударений у слов. Особенно это важно при анализе стихов. Однако стандартные алгоритмы акцентуации проставляют ударения на основе современных норм русского языка, в то время как ритмический рисунок и рифма в стихотворениях многих авторов работает только при той орфоэпической норме, которая была принята в тот временной период развития русского языка, когда они создавали свои произведения. Таким образом, мы видим своем задачей разработку алгоритмов для автоматической расстановки ударений, который проводит акцентуацию с учетом норм, которые были использованы автором при написании своего произведения. В данной работе будут представлены два алгоритма, решающие эту задачу, имитируя нормы, которые использовал А. С. Пушкин.

Исходный набор данных был обработан двумя способами для тестирования каждого из предложенных в статье алгоритмов. Для первого алгоритма, который использовался в качестве базового метода, был создан специальный корпус, основанный на первых четырёх томах собрания сочинений А. С. Пушкина в десяти томах. Во всех стихотворениях были проставлены ударения с помощью акцентора, основанного на грамматическом словаре русского языка А. А. Зализняка. Обработанные стихотворения затем преобразовывались в пары последовательностей, состоящие из входной последовательности букв и желаемой выходной последовательности нулей и единиц, причём единица для элемента желаемой выходной последовательности означает то, что соответствующий этому элементу символ входной последовательности должен находиться под ударением, а ноль – что нет. Первая из последовательностей поступала на вход алгоритму, который делал ее предобработку перед использованием в качестве обучающего множества для нейросети. Каждый символ получал свой набор признаков: является ли он заглавной/строчной буквой, буквой/цифрой/знаком препинания, предыдущая перед текущим символом буква, следующая за текущим символом буква.

Для другого варианта алгоритма расстановки ударений, основанного на проставлении ударения не для целого текста, а для одного слова, использо-

* Работа осуществлена в рамках гранта РНФ № 19-18-00466 «Разработка и реализация информационной системы многоуровнего исследования стихотворных текстов»

вался другой способ представления данных. Здесь на вход подается цепочка символов слова, а также его морфологическая характеристика, что позволяет автоматически разрешать неодзнозначность в омографах. При обучении второго алгоритма использовался метод переноса обучения (transfer learning), поэтому для первичного обучения также был использован корпус прозаических текстов, в которых были расставлены ударения. Количество уникальных слов в обучающей выборке составило более 800 тысяч.Дообучение проводилось с помощью второго массива данных – набора слов из Конкорданса стихов А. С. Пушкина Дж. Т. Шоу.

При работе с базовым методом обработанные первым способом последовательности поступали на вход алгоритма, использующего метод условных случайных полей (Conditional random fields, CRF).

Второй алгоритм использует для решения задачи классификации на ударные и безударные позиции рекуррентные нейронные сети типа GRU (Gated recurrent unit) с применением методики переноса обучения.

В результате экспериментов было показано, что вариант алгоритма на основе рекуррентных нейронных сетей типа GRU, инициализируемых состоянием на основе векторного представления морфохарактеристик слова, обеспечивает 5,6% ошибок в расстановке ударений. Базовый алгоритм, основанный на условных случайных полях с использованием признаков символов, дает 15% ошибок на тестовом множестве.

Author Index

Научное издание

Data Analytics and Management in Data Intensive Domains Extended Abstracts of the XXII International Conference DAMDID / RCDL'2020

October 13–16, 2020
Voronezh, Russia

Минимальные системные требования:
PC не ниже класса Pentium I, 32 Mb RAM,
свободное место на HDD 16 Mb,
Windows 95/98, Adobe Acrobat Reader,
дисковод CD-ROM 2-х, мышь

Подписано к использованию 12.10.2020
Объем данных 9 Мб. 1 электрон. опт диск (CD-ROM).
Тираж 500 экз. Заказ 200

ООО «ВЭЛБОРН»
Издательство «Научно-исследовательские публикации»
394068, г. Воронеж, Московский пр-т, 98
Тел. +7 (930) 403-54-18
http://www.scirep.ru E-mail: publish@scirep.ru

Изготовлено фирмой «Большой формат»
(ООО «Твой выбор»)
394018, г. Воронеж, ул. Кости Стрелюка, д. 11/13, офис 6
Тел. +7 (473) 238-26-38
http://big-format.ru, E-mail: 382638@mail.ru