National University of Science and Technology

NUST MISIS

Data Analytics and Management in Data Intensive Domains

Book of Abstracts

of the XXIII International Conference DAMDID / RCDL'2021

October 26 – 29, 2021

Moscow, Russia

Edited by Alexandra Khvan, Vladimir Cheverikin,

Semen Kuzovchikov

Moscow 2021

УДК [004.6:004.8]

ISBN 978-5-907227-93-4

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 trans-formation of the RCDL conference (“Digital libraries: advanced methods and technologies, digital collections”, http://rcdl.ru) so that the continuity with RCDL has been preserved after many years of its successful work.

ISBN 978-5-907227-93-4

CONTENT

Machine Learning for Materials Science

Machine Learning Accelerated Multicomponent Alloy Design

Yi Liu[1]
[1]Shanghai University
*yiliu@shu.edu.cn**

The core of Materials Genome approach is the combination of high-throughput computation, high-throughput experiment, and data analysis (machine learning), aiming to accelerate materials design and development at lower cost. In this talk we will show two data-driven studies in multicomponent alloy design based on computation and experimental data, respectively, as follows:

(1) Computation & Machine Learning: In alloy design, it is fundamentally important to clarify the preferential site occupancy of alloying elements to elaborate the strengthening mechanisms. It is, however, a formidable task for first-principles (FP) calculations to explore the enormous potential doping configurations in the complex multi-component alloys. In this work, we first carried out high-throughput FP calculations systematically for several hundred alloy doping configurations in superalloy, considering ~10 alloying element substitution at multiple nonequivalent sites. The machine learning models can be used more efficiently for further prediction, reducing the cost of expensive FP calculations while maintaining the certain accuracy. We developed the machine learning models based on the high-throughput FP calculated data. Specifically, we designed a “Center Environment” (CE) feature model to construct descriptive features by combining elemental properties and local composition and structure information of both center and environment. It is shown that the CE descriptors can be used to predict both the substitution energy and local geometry of alloying elements in superalloy. By comparison we show clearly that the machine learning prediction using feature construction with both composition and structure information is more accurate and robust than that with composition only. Taking the advantages of the accuracy of first-principle calculations and efficiency of machine learning methods, such combined FP-ML approach becomes an emerging strategy to explore enormous configurations commonly required in computational materials simulations and design.

(2) Experiment & Machine Learning: Conventional trail-and-error experiment development approaches rely heavily on the intuition and experience of researchers and are limited by the low efficiency of single sample experiment mode. To accelerate materials discovery, we developed a machine learning (ML) aided high-throughput experiment (HTE) approach to

optimize the composition of non-equimolar hard high entropy alloy (HEA) CoxCryTizMouWv. The HTE approach conduct experiments at a batch sample mode featuring mutli-station, automation, and parallelization, more efficient than the conventional single sample experiment mode. We designed a set of HTE facilities covering whole preparation process of bulk alloy from multi-station ingredient assignment to multi-station sample synthesis and specimen preparation. The HTE approach improves the efficiency but it is still hard to explore all potential combinatorial compositions. To further accelerate the development process, we designed and conducted only the fractional HTEs with the aid of the ML prediction. First we designed initial 111 experiments by varying Mo and W compositions critical to hardness as well as important descriptors related to phase structures, considering both the importance and the diversity of data. At the basis of these preliminary experiment data, we then developed 12 ML methods combining three ML algorithms and four groups of descriptors. After cross-validation and independent data evaluation on 120 ML models, the preliminary good ML models were selected to design 27 compositions covering various hardness ranges. The final ML models were obtained using the all 138 experiment data and predicted the hardness with the mean relative errors of ~6% and ~15% at the high/medium (HV> 600) and low (HV< 600) hardness ranges, respectively. To extract knowledge out of machine learning, the multiple ML models were used to predict the hardness of hypothetical 3876 alloys covering the whole composition range, showing the consistent trends of component effects from the complete composition-hardness and descriptor-hardness correlations. The hardening mechanisms were finally discussed based on the microstructures of equimolar CoCrTiMoW. This work proved that the machine learning guided high-throughput experiment (ML-HTE) approach becomes an effective strategy for multicomponent alloy development with possible hundred times overall acceleration at lower cost.

Machine Learning Application to Predict New Inorganic Compounds – Results and Perspectives

Nadezhda N. Kiselyova [1]; Victor A Dudarev [2]; Andrey Stolyarenko [1]
[1]IMET RAS
[2]NRU HSE
kis@imet.ac.ru; vic_dudarev@mail.ru; stol-drew@yandex.ru

A brief overview of the problems is given in the field of inorganic chemistry and materials science, solved using machine learning (ML). The main ML methods limitations and the subject area peculiarities are considered that must be taken into account when using ML. Solved problems examples of new inorganic com-pounds design and the results of comparing predictions

with new experimental data are given. Systems developed by the authors are considered that aimed to not yet obtained inorganic compounds design, based on MO methods, as well as promising directions for such systems development in order to improve the pre-dictions accuracy for new substances and their corresponding properties values estimations.

Artificial intelligence, machine learning and data-enabled multiscale simulation workflows for the design and development of molecular materials

Matteo Baldoni[1] ; Fabio Le Piane[1]; Francesco Mercuri[1]
[1]ISMN-CNR
matteo.baldoni@ismn.cnr.it; fabio.lepiane@ismn.cnr.it; francesco.mercuri@cnr.it

Machine Learning tools are nowadays widely applied extensively to the prediction of the properties of molecular materials, using datasets extracted from high-throughput computational models. In several cases of scientific and technological relevance, the properties of molecular materials are related to the link between molecular structure and phenomena occurring across a wide set of spatial scales, from the nanoscale to the macroscale. Here, we describe an approach for predicting the properties of molecular aggregates based on multiscale simulations and machine learning.

Keywords: Computational modelling · Machine learning · Ontology develop- ment · Data science.

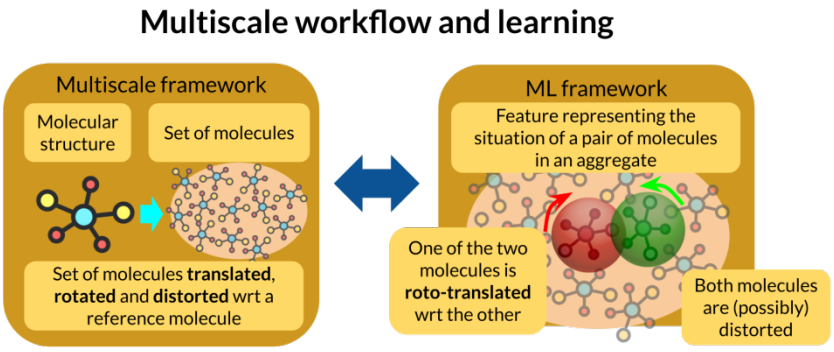

Fig. 1. Connection between the multiscale framework, for the evaluation of the mor- phology in molecular aggregates, and the ML framework, defining features that describe molecular pairs in aggregates.

In recent years, machine learning (ML) methods have applied with success to studies of the properties of molecular materials[2].The vast majority of these studies are focused on the properties of individual molecules, targeting the correlation between molecular structure and resulting properties. The properties of several technological materials constituted by molecular aggregates, however, depend on both molecular structure and on aggregation morphology, as for exam- ple in the case of nanoscale materials[4]. Computational methods for predicting the properties of molecular materials must therefore integrate the properties of individual molecules with information about aggregation morphology, which, in turn, can be related to materials fabrication and processing[3]. The definition of a modelling paradigm able to simulate and predict the properties of molecular materials as a function of molecular structure and aggregation/fabrication conditions can potentially enable high-throughput development of novel materials for technological applications. In this work, we design and implement a computational workflow for the simulation of the properties of molecular materials integrated with a ML model for enhancing the computational workload. The workflow is based on a multi-scale top-down approach, in which target properties are defined from the application to the molecular scale. The workflow is implemented through top-down hierarchical data structures, which connects the properties of molecular materials at the nanoscale to the atomistic/electronic scale. Modelling data are generated by applying domain-specific simulation protocols based on atomistic molecular dynamics (MD)[5] and density functional theory (DFT) calculations[1]. ML approaches are therefore applied to enable the scale reduction, providing a local mapping at a lower scale of the properties of large molecular aggregates, reducing greatly the overall computational load.

We then proceeded implementing a ML algorithm to test the potential of our approach. In particular, we aimed to find a more efficient way to calculate the electronic coupling between two molecules. We implemented both a random forest regressor and a KRR algorithm, achieving a good overall accuracy. However, as already pointed out in literature, in this kind of applications the development of a good representation of a molecule that can be easily ingested by the ML algorithm proved to be a more critical task than the development of the actual ML algorithm. We found that a representation that includes orientation, distance and the amount of molecular distortion compared to an idealized molecule (RMSD) proved to be a functional representation of our system. In the end, we achieved a system that, once trained, could predict our target property nearly instantly on a standard desktop CPU, while a standard DFT calculation requires almost 40 minutes for each pair on an HPC facility.

Ongoing work is focused on increasing the predictive performance of our model, while also improving the expressiveness of our representation. In particular, we are testing more advanced Deep Learning (DL) algorithms, which in turn usually requires more data to perform adequately.

References

Baldoni, M., Lorenzoni, A., Pecchia, A., Mercuri, F.: Spatial and ori- entational dependence of electron transfer parameters in aggregates of iridium-containing host materials for OLEDs: Coupling constrained den- sity functional theory with molecular dynamics. Physical Chemistry Chem- ical Physics 20(45), 28393–28399 (2018). https://doi.org/10.1039/c8cp04618b, https://pubs.rsc.org/en/content/articlehtml/2018/cp/c8cp04618b

Butler, K.T., Davies, D.W., Cartwright, H., Isayev, O., Walsh, A.: Machine learning for molecular and materials science. Nature 559(7715), 547–555 (2018). https://doi.org/10.1038/s41586-018-0337-2

Le Piane, F., Baldoni, M., Mercuri, F.: Predicting the properties of molecular mate- rials: Multiscale simulation workflows meet machine learning. arXiv pp. 1–14 (2020)

Liu, H., Xu, J., Li, Y., Li, Y.: Aggregate nanostructures of organic molecular materials. Accounts of Chemical Research 43(12) (2010). https://doi.org/10.1021/ar100084y

Lorenzoni, A., Mosca Conte, A., Pecchia, A., Mercuri, F.: Nanoscale mor- phology and electronic coupling at the interface between indium tin ox- ide and organic molecular materials. Nanoscale 10(19), 9376–9385 (2018). https://doi.org/10.1039/c8nr02341g, http://dx.doi.org/10.1039/c8nr02341g

Simulations of mechanical properties of materials by using machine learning interatomic potentials

Faridun Jalolov[1] ; Artem Oganov[1] ; Alexander G Kvashnin[1]
[1]Skolkovo Institute of Science and Technology
Faridun.Jalolov@skoltech.ru; a.oganov@skoltech.ru; A.Kvashnin@skoltech.ru

Mechanical properties of material are the main characteristics which determine suitability of material to be or not to be applied in industrial applications. Nowadays elastic and mechanical characteristics of single crystals can be calculated by using density functional theory with high accuracy compared to experimental data. Moving towards more complex materials (polycrystals, heterostructures, composites etc.) it becomes more and more complicated to accurately calculate mechanical characteristics. One problem is the size limitation. Due to large sizes of polycrystals or composites it is not possible to use accurate DFT methods for such calculations. From other side, empirical potentials can easily handle such calculations, but the accuracy will be quite low. Moreover, new atomic configurations and new unexpected compositions often cannot be accurately calculated by empirical potentials. In this situation, machine-learning interatomic potentials (MLIPs) have direct indications for use.

Here we use moment tensor potentials (MTP) to solve this issue [1,2]. These models are trained on the results of quantum mechanical calculations and reproduce the behavior of ab-initio models after that. Performance of MLIPs is comparable with empirical potentials, and the accuracy is close to that of quantum-mechanical models.

Thus, we start our study with development of special code allowing the calculations of elastic tensor of selected materials by using MTP with active learning technique]2]. We performed the calculations of elastic tensor, bulk and shear moduli of several known covalent materials, namely diamond, SiC etc, compare obtained data with results of DFT calculations and available experimental data.

References

[1] A. Shapeev, Multiscale Model. Simul. 14, 1153 (2016).

[2] E.V. Podryabinkin and A.V. Shapeev, Computational Materials Science 140, 171 (2017).

Fast predictions of lattice energies by continuous isometry invariants of crystal structures

Jakob Ropers[1] ; Marco Mosca[1] ; Olga Anosova[1]; Vitaliy Kurlin[1]; Andrew Cooper[1]
[1]University of Liverpool
jakob.ropers@theropers.org; m.m.mosca@liverpool.ac.uk; oanosova@liverpool.ac.uk; vkurlin@liv.ac.uk; aicooper@liverpool.ac.uk

Crystal Structure Prediction (CSP) aims to discover solid crystalline materials by optimizing periodic arrangements of atoms, ions or molecules. CSP takes weeks of supercomputer time because of slow energy minimizations for millions of simulated crystals. The lattice energy is a key physical property, which hints at thermodynamic stability of a crystal but has no simple analytic expression. Past machine learning approaches to predict the lattice energy used slow crystal descriptors depending on manually chosen parameters. The new area of Periodic Geometry offers much faster isometry invariants that are also continuous under perturbations of atoms. Our experiments on simulated crystals confirm that a small distance between the new invariants guarantees a small difference of energies. We compare several kernel methods for invariant-based predictions of energy and achieve the mean absolute error of less than 5kJ/mole or 0.05eV/atom on a dataset of 5679 crystals.

Databases on Properties of Substances and Materials

Implementation of Materials Data Integration using Ontology

Toshihiro Ashino[1]; Nobutaka Nishikawa[2]; Masahiko Demura[1]; Takuya Kadohira[1]
[1]National Institute for Materials Science
[2]Mizuho Research & Technologies, Ltd.
ashino@acm.org; nobutaka.nishikawa@mizuho-ir.co.jp; DEMURA.Masahiko@nims.go.jp; KADOHIRA.Takuya@nims.go.jp

For materials design, it is necessary to refer a variety of heterogeneous in-formation resources, such as databases, formulae, and computational simulations. Ontology-based data integration is one of the approaches to combine data from multiple heterogeneous data resources. In this paper, an implementation of ontology-based data integration with Semantic Web standards and its application for materials data integration are presented.

Molecular and Materials Basic Ontology: development and first steps

Fabio Le Piane[1]; Matteo Baldoni[1]; Mauro Gaspari[2]; Francesco Mercuri[1]
[1]ISMN-CNR
[2]University of Bologna
fabio.lepiane@ismn.cnr.it; matteo.baldoni@ismn.cnr.it; mauro.gaspari@unibo.it; francesco.mercuri@cnr.it

Advanced materials and their applications have become a key field of research, and it looks like this trend is not going to change soon. For that reason, the need for systematic and efficient methods for organizing knowledge in the field and conduct computational or experimental investigations is stronger than ever.

In this work, we present a basic implementation of MAMBO - an ontology for molecular materials and their applications in real-life scenarios. The development of MAMBO has been guided by the needs of the research community involved in the development of novel materials with functional properties, with particular attention to the nanoscale. MAMBO aims at extending the current work in the field, while retaining a modular nature in order to allow straightforward extension of concepts and relations to neighboring domains.

Our work is expected to enable the systematic integration of computational and experimental data in specific domains of interest (nanomaterials, molecular materials, organic an polymeric materials, supramolecular and bio-organic systems, etc.). Moreover, MAMBO is developed with a strong focus on the applications of data-driven frameworks for the design of novel materials with tailored characteristics.

Interoperability and architecture requirements analysis and metadata standardization for a research data infrastructure in catalysis

Martin T Horsch[1]; Taras Petrenko[1]; Volodymyr Kushnarenko[1]; Björn Schembera[1]; Bianca Wentzel[2]; Alexander Behr[3]; Norbert Kockmann[3]; Sonja Schimmler[2]; Thomas Bönisch[1]
[1]High Performance Computing Center Stuttgart
[2]Fraunhofer Institute for Open Communication Systems
[3]TU Dortmund
martin.horsch@hlrs.de; taras.petrenko@hlrs.de; volodymyr.kushnarenko@hlrs.de; bjoern.schembera@hlrs.de; bianca.wentzel@fokus.fraunhofer.de; alexander.behr@tu-dortmund.de; norbert.kockmann@tu-dortmund.de; sonja.schimmler@fokus.fraunhofer.de; thomas.boenisch@hlrs.de

The National Research Data Infrastructure for Catalysis-Related Sciences (NFDI4Cat) is one of the disciplinary consortia formed within the German national research data infrastructure (NFDI), an effort undertaken by the German federal and state governments to advance the digitalization of all scientific research data within the German academic system in accordance with the FAIR principles. This work reports on initial outcomes from the NFDI4Cat project. The data value chain in catalysis research is analysed, and architecture and interoperability requirements are identified by conducting user interviews, collecting competency questions, and exploring the landscape of semantic artefacts. Methods from agile software development are employed to collect, organize, and present the collected requirements; workflows are annotated on the basis of metadata standards for research data provenance, by which requirements for domain ontologies in catalysis are deduced.

Searching for an Optimal Data Platform for Relevant Information Search in Inorganic Chemistry and Materials Science

Victor A Dudarev[1]; Sergey Babikov[1]
[1]NRU HSE
vic_dudarev@mail.ru; ssbabikov@edu.hse.ru

Choosing the most suitable database management system is one of the most critical challenges in developing any information system operating on big data. When selecting, as a rule, the overall system speed is considered the main criterion regarding certain data structures due to the subject area specifics. In the current article, using the example of searching for relevant information on inorganic com-pounds, an attempt is made to analyze the possibility of using relational and graph database management systems (DBMSs) to build a data storage subsystem. Graph-based database implementations, powered by SQL Graph and Neo4j, are considered and compared with a relational version based on SQL Server. Typical query execution speed comparative analysis is carried out when searching for relevant information in the field of inorganic chemistry and materials science.

Metadata Schema to support FAIR Data in Scanning Electron Microscopy

Reetu Elza Joseph[1]; Aditya Chauhan[1]; Catriona Eschke[2]; Ahmad Zainul Ihsan[3]; Mehrdad Jalal[1]; Ute Jäntsch[1]; Nicole Jung[1]; C. N. Shyam Kumar[4]; Christian Kübel[1]; Christian Lucas[2]; Matthias Mail[1]; Andrey Mazilkin[1]; Charlotte Neidiger[1]; Mirco Panighel[5]; Stefan Sandfeld[3]; Rainer Stotzka[1]; Richard Thelen[1]; Rossella Aversa[1]

[1]Karlsruhe Institute of Technology

[2]Helmholtz-Zentrum Hereon

[3]Forschungszentrum Jülich

[4]CEA Saclay

[5]Consiglio Nazionale delle Ricerche - Istituto Officina dei Materiali

*reetu.joseph@kit.edu; aditya.chauhan@kit.edu; catriona.eschke@hereon.de; a.ihsan@fz-juelich.de; mehrdad.jalali@kit.edu; ute.jaentsch@kit.edu; nicole.jung@kit.edu; shyam-kumar.chethalaneelakandhan@cea.fr; christian.kuebel@kit.edu; christian.lucas@hereon.de; matthias.mail@kit.edu; andrey.mazilkin@kit.edu; charlotte.neidiger@kit.edu; panighel@iom.cnr.it; s.sandfeld@fz-juelich.de; rainer.stotzka@kit.edu; richard.thelen@kit.edu; rossella.aversa@kit.edu**

The development and the adoption of metadata schemas and standards are a key aspect in data management. In this paper, we introduce our approach to a metadata model in the field of Materials Science. We present the specific use case of a metadata schema for Scanning Electron Microscopy, a characterization technique which is routinely used in Materials Science. This metadata schema is aiming to be a de-facto standard which will be openly available for reuse and further extension to other electron microscopy techniques. The development and the adoption of metadata schemas and standards are a key aspect in data management. In this paper, we introduce our approach to a metadata model in the field of Materials Science. We present the specific use case of a metadata schema for Scanning Electron Microscopy, a characterization technique which is routinely used in Materials Science. This metadata schema is aiming to be a de-facto standard which will be openly available for reuse and further extension to other electron microscopy techniques.

Computational Materials Science

Computational prediction, Synthesis, Properties, and Crystal Structure of WB5-x

Alexander G Kvashnin[1]; Dmitry Rybkovskiy[1]; Vladimir Filonenko[2]; Vadim Brazhkin[2]; Artem Oganov[1]
[1]Skolkovo Institute of Science and Technology
[2]High Pressure Physics Institute
A.Kvashnin@skoltech.ru; D.Rybkovskiy@skoltech.ru; vpfil@mail.ru; brazhkin@hppi.troitsk.ru; a.oganov@skoltech.ru

Operation of many of the industrial application is not possible without using superhard materials. Nowadays it is important to search for new cheap and effective materials which can substitute traditional materials in many field of science and technology. Traditionally material can be called as superhard if its Vickers hardness is higher than 40 GPa [1–3]. Here we predict new tungsten and molybdenum borides, some of which are promising hard materials that are expected to be thermodynamically stable in a wide range of conditions. We computed the composition-temperature phase diagram, which shows the stability ranges of all predicted phases. New boron-rich compound WB5 is predicted to be superhard with Vickers hardness of 45 GPa [4]. We performed the experimental synthesis and structural description of a boron-rich tungsten boride and measurements of its mechanical properties are performed [5]. The ab initio calculations of the structural energies corresponding to different local structures make it possible to formulate the rules determining the likely local motifs in the disordered versions of the WB5 structure, all of which involve boron deficit. The generated disordered WB4.18 and WB4.86 models both perfectly match the experimental data, but the former is the most energetically preferable. The precise crystal structure, elastic constants, hardness, and fracture toughness of this phase are calculated, and these results agree with experiment. Mild synthesis conditions (enabling a scalable synthesis) and excellent mechanical properties make WB5-x a very promising material for the drilling technology.

Reference

[1] V. L. Solozhenko, S. N. Dub, and N. V. Novikov, Mechanical Properties of Cubic BC2N, a New Superhard Phase, Diam. Relat. Mater. 10, 2228 (2001).

[2] V. L. Solozhenko and E. Gregoryanz, Synthesis of Superhard Materials, Mater. Today 8, 44 (2005).

[3] V. L. Solozhenko, O. O. Kurakevych, D. Andrault, Y. Le Godec, and M. Mezouar, Ultimate Metastable Solubility of Boron in Diamond: Synthesis of Superhard Diamondlike BC5, Phys. Rev. Lett. 102, 015506 (2009).

[4] A. G. Kvashnin, H. A. Zakaryan, C. Zhao, Y. Duan, Y. A. Kvashnina, C. Xie, H. Dong, and A. R. Oganov, New Tungsten Borides, Their Stability and Outstanding Mechanical Properties, J. Phys. Chem. Lett. 9, 3470 (2018).

[5] A. G. Kvashnin, D. V. Rybkovskiy, V. P. Filonenko, V. I. Bugakov, I. P. Zibrov, V. V. Brazhkin, A. R. Oganov, A. A. Osiptsov, and A. Ya. Zakirov, WB5–x: Synthesis, Properties, and Crystal Structure. New Insights Into the Long-Debated Compound, Adv Sci 7, 200775 (2020).

Usage of Robust Regression for Approximation of Thermodynamic Data

Alexey L Voskov[1]
[1]Lomonosov Moscow State University
alvoskov@gmail.com

M-estimators based on Huber and Andrews sine loss functions were successfully used for approximaton of heat capacities and heat contents of K-substituted natrolite and petalite by means of the weighted sum of Einstein functions. It automatically excluded outliers for petalite and narrow peak of lambda-transition for K-natrolite.

AICON2: A program for calculating transport properties quickly and accurately

Tao Fan[1]
[1]Skolkovo Institute of Science and Technology
*Tao.Fan@skoltech.ru**

Calculating the transport properties, such as electrical conductivity, has been a great challenge in materials modeling fields because of its complexity. We have implemented an algorithm to calculate the electrical transport properties using the generalized Kane band model and perturbation theory in the framework of the relaxation time approximation. Three scattering mechanisms affect the total relaxation time: acoustic phonon scattering, polar optical phonon scattering, and ionized impurity scattering. All the necessary parameters can be calculated from first principles. Moreover, the program can also calculate the lattice thermal conductivity based on a modified Debye-Callaway model and all the input parameters are from first principles calculations. The capability of the program was tested on a group of semiconductors, and the obtained results show reasonable agreement with experiment. The program works fast, and is robust and especially appropriate for high-throughput screening of thermoelectric materials.

Conceptual modeling, Data Integration, Ontologies and Applications

A Needs-Based Augmented Reality System

Manal A Yahya[1]; Ajantha Dahanayake[1]
[1]Lappeenranta University of Technology
manal.yahya@student.lut.fi; ajantha.dahanayake@lut.fi

Augmented reality aims to enhance the real world with computer-generated information. AR technology is both attractive and promising. Current AR experiences depend on external elements to launch, such as markers, images, and location. For an AR experience to be more personalized, this research proposes a scheme to trigger AR experiences based on human needs. This approach should enable capturing human needs, analyzing them to select the most suited experiences that fulfill or aids in fulfilling needs. The contribution of this paper includes (1) a study of current AR technologies and triggers, (2) an analysis of human needs into measurable elements (3) a description of a needs-based AR application process.

Response to cybersecurity threats of informational infrastructure based on conceptual models

Nikolay Kalinin[1]; Nikolay A. Skvortsov[2]
[1]MSU, Faculty of Computational Mathematics and Cybernetics
[2]FRC CSC RAS
Kalinin-NA@yandex.ru; nskv@mail.ru

Response to the threats of information security in conditions of modern organization with a large infrastructure is an area with emergency loaded intensity of the data usage. For a successful exposure and the prevention of computer attacks the construction of complex models of the events and infrastructure is required. In this work, the question of the applicability of ontological models and reasoning for supporting response process is examined. On the basis of built ontology, practical use cases are demonstrated.

Prospects for the evolution of the Russian segment of the Virtual Atomic and Molecular Data Center (VAMDC)

Alexey Akhlyostin[1]; Nikolay Lavrentiev[1]; Alexey Privezetsev[1]; Alexander Z. Fazliev[1]
[1]Institute of Atmospheric Optics SB RAS
lexa@iao.ru; lnick@iao.ru; remake@iao.ru; fazliev@yandex.ru

The report discusses the prospects for the development of Russian sites, which accumulate atomic, ionic and molecular spectral data included in the European Virtual Center for Atomic and Molecular Data. The key issues in the development of the segment are the semantisation of tabular and graphical information resources and the personalization of data and their properties. It is shown how

the issues of constructing by researchers their own expert arrays based on the knowledge base of spectral information resources and physical quantities, as well as the actualization of expert data.

On a Conceptual Data Model with Orientation to Data Integration

Manuk Garush Manukyan[1]
[1]Yerevan State University
mgmanukyan@gmail.com

In this paper a conceptual data model oriented to data integration is proposed. Formal definition of the considered conceptual data model is provided. To define the behavior of entities of the conceptual level, an algebra over such entities was developed. Formalization issues of data integration concept are discussed. Principles of mapping of source data models basic constructions into conceptual data model are considered. Mapping from data sources into conceptual schema is defined as an algebraic program.

Towards a Recommender System for the Choice of the Universal Decimal Classification Code for Mathematical Articles

Olga A. Nevzorova[1]; Damir Almukhametov[1]
[1]Kazan Federal University
onevzoro@gmail.com; dnlanik@gmail.com

Authors of scientific papers in the field of mathematics usually use the universal decimal classification scheme to search for related articles. UDC is a hierarchical classification scheme that allows librarians and editors to specify one or more codes for publications. Typically, the classification code identifies a subject editor who is responsible for the review process for articles submitted to scientific journals. In this article, we will explore a new approach to assigning UDC code for mathematical work, based on the OntoMathPRO ontology.

This ontology is an applied ontology for the automatic processing of professional mathematical articles in Russian and English. An ontology defines concepts commonly used in mathematics, as well as an evolving and poorly established vocabulary extracted from contemporary scientific articles. OntoMathPRO covers a wide range of areas of mathematics such as number theory, set theory, algebra, analysis, geometry, computation theory, differential equations, numerical analysis, probability theory, and statistics. Each class has a textual explanation, Russian and English inscriptions, including synonyms.

We investigated a set of classification functions, which are presented as ontology concepts, and identified the most relevant ones for constructing code maps of some UDC codes in the field of

mathematics. We found that the code maps of the considered UDC codes can be built on the basis of the selected features (method, equation, problem). The values of these features are determined using the OntoMathPRO ontology. The constructed code maps allow for successfully assigning the considered UDC codes for publications.

Problem Solving Architectures and Infrastructures, Experiment Organization

A Survey of Big Data Pipeline Orchestration Tools from the Perspective of the DataCloud Project

Mihhail Matskin[1]; Shirin Tahmasebi[1]; Amirhossein Layegh[1]; Amir H. Payberah[1]; Aleena Thomas[2]; Nikolay Nikolov[2]; Dumitru Roman[2]
[1]KTH Royal Institute of Technology, Computer Science
[2]SINTEF
misha@kth.se; shirin_tahmasebi94@yahoo.com; amir.layegh1994@gmail.com; payberah@kth.se; Aleena.Thomas@sintef.no; nikolay.nikolov@sintef.no; Dumitru.Roman@sintef.no

This paper presents a comparative survey of existing solutions for Big Data pipeline orchestration based on a comparative framework developed in the DataCloud project. We propose criteria for evaluating the solutions to support reusability, flexible communication modes, and separation of concerns in pipeline descriptions. This survey aims to identify research and technological gaps and to recommend approaches for filling the identified gaps. Further work on the project will be oriented towards the design, implementation, and practical evaluation of the recommended approaches.

A FAIR Problem-Solving Lifecycle Architecture

Nikolay A. Skvortsov[1]
[1]FRC CSC RAS
nskv@mail.ru

Research infrastructures are intended to provide access to scientific data and resources needed for problem-solving. Approaches to ensuring interoperability and reuse of heterogeneous resources in such infrastructures are necessary investigations. Researchers tend to integrate and reuse existing resources but do not spend much effort to publish the resources created during solving scientific problems to make them reusable in communities. As the result, further users of these results have to spend their time and effort on integrating resources to reuse them. We propose an architecture of research infrastructures that are initially based on the lifecycle of problem-solving in research communities providing interoperability and reuse of the sources and results of problem-solving in research communities. In this architecture, most of the maintenance tasks are moved from the data integration stage to the data publication stage, and data are manipulated in accordance with the domain specifications accepted by communities. This makes providing the interoperability and reuse of resources in the infrastructure more competent and trivial.

Comparison of Data-Driven Approaches to Modeling Complex Behavior of 2D Liquid Simulator

Dmitry Kovalev[1]; Dmitry Khliustov[2]; Sergey Safonov[1]
[1]Aramco Research Center, Moscow, Aramco Innovations LLC
[2]Lomonosov Moscow State University
dmitry.kovalev@aramcoinnovations.com; hlustov.d@gmail.com;
sergey.safonov@aramcoinnovations.com

Modeling complex system dynamics traditionally is implemented with the use of differential equations, which requires hand-crafted work of a qualified expert and significant amount of time. The advent of data-driven approaches allows to overcome these difficulties and substitute traditional models with models built in automated way directly from observations. This paper compares several data-driven approaches to modeling 2D liquid simulator. Dataset is generated from it for both training and testing with fixed simulator parameters. Local and global types of models are evaluated with metrics, describing different aspects of liquid behavior (spatial, spatio-temporal and worst-case settings). Other metrics introduced allow to capture differences not only in distances, but also in distributions, which is more natural for human perception and enables to quantitively compare similar pictures. From the model evaluation, it is inferred that the use of decomposition improves overall accuracy and the trajectories figures, though at the same time model generalizability decreases. On the other hand, utilizing locality leads to more generalizable models at the cost of accuracy. Model training and inference times are provided and main directions for further research are outlined.

MLDev: Data Science Experiment Automation and Reproducibility Software

Anton Khritankov[1]; Nikita Pershin[1]; Nikita Ukhov[1]; Artem Ukhov[1]
[1]MIPT
anton.khritankov@phystech.edu; pershin.nv@phystech.edu; uhov.aa@phystech.edu;
uhov.na@phystech.edu

In this paper, we explore the challenges of automating experiments in data science. We propose an extensible experiment model as a foundation for integration of different open source tools for running research experiments. We implement our approach in a prototype open source MLdev software package and evaluate it in a series of experiments yielding promising results. Comparison with other state-of-the-art tools signifies novelty of our approach.

Machine Learning Applications

Predicting the increase in postoperative motor deficits in patients with supratentorial gliomas using machine learning methods

Eugene A Ilyushin[1]
[1]Lomonosov Moscow State University
e.ilyushin@cs.msu.ru

Surgery of glial tumors of the brain located in the motor areas vicinity is associated with a high risk of increasing neurological deficits. Motor deficit affects overall survival in this group of patients. Nowadays, no method allows for an objective preoperative data-based prognosis of the risk of neurological impairment in each particular case. Objective: develop a convolution neuronal network that can predict motor worsening in patients with supranational gliomas using the preoperative MRI data. Materials and methods: the study included 527 patients aged 18 years and older with newly identified supratentorial gliomas. All patients underwent preoperative MRI and tumor removal based on Burdenko National Center of neurosurgery in 2013-2019. Data on motor status dynamics after surgery for these patients were obtained from the electronic medical records using the original semiautomatic algorithm for natural language processing. The T2FLAIR mode is used for training our model. The model demonstrates the following metrics of quality: accuracy 91\%, sensitivity 94%, specificity 89%, ROC AUC 91%, and F1 92\%. Thus, machine learning methods predict the motor worsening with relatively high accuracy in patients with supratentorial gliomas at the preoperative stage, based on brain MRI data.

Social Network Analysis of the Professional Community Interaction - Movie Industry Case

Ilia Andreevich Karpov[1]; Roman Marakulin[1]
[1]International Laboratory for Applied Network Research, NRU HSE
karpovilia@gmail.com; marakromrom@gmail.com

With the rise of the competition in the movie production market, because of the new players, such as Netflix, Hulu, HBO Max, and Amazon Prime, that are aimed on producing a large amount of exclusive content in order to get competitive advantage, it is extremely important to minimize the number of unsuccessful titles. This paper tries to analyze the movie industry community by collecting unique

data, creating the actor-casting director-talent agent-director graph and applying social network analysis approaches. Also, the information obtained during the analysis is used to improve the movie success prediction models.

The Analysis of Trajectories in Moscow Subway

Mariia Nekraplonna[1]; Dmitry Namiot[1]
[1]Lomonosov Moscow State University
maria.nekraplennaya@gmail.com; dnamiot@gmail.com

Along with the continuous growth of megacities, their transportation systems have become increasingly large and complex. The use of transportation systems by passengers directly reflects the changes that occur in the urban environment - for this reason, the study of urban mobility is an important task of digital urbanism. In particular, this paper is devoted to the study of spatial patterns (repetitive routes) in transportation systems with the case study on the Moscow subway. A brief review of data mining approaches to transportation systems data in general and to the task of spatial patterns extraction, in particular, is presented. A simple method for pattern extraction is proposed and applied to the

Moscow subway data. As a result of the deployment of the proposed method the list of patterns was obtained - the graph of spatial patterns of the transport system under study was constructed based on it.

Multi-Modal Human Cognitive State Recognition during Reading

Nikita Filimonov[1]
[1]Moscow State University
*marxk1178@gmail.com**

Human cognitive state recognition is an important and chal- lenging task. Various registration technologies can be used to collect physiological data that potentially contains relevant information regard- ing current cognitive state of a human subject. Oculography (eye-tracking) and electroencephalography (EEG) are most popular and well-researched registration technologies, both technologies have cheap commercial vari- ants that do not require laboratory equipment or involvement of profes- sional physiologist to collect the data. However, it is still problematic and expensive to obtain large-scale datasets of physiological data of such sort. In this work a review and analysis of available open source phys- iological data is provided, a task of natural reading is considered since work of eyes and human brain during reading are of great interest for cognitive psychology. A multi-modal approach that involves combining EEG and eye-tracking data in jointly trained artificial neural network is proposed. Intermediate results are presented regarding encoding EEG signals with Variational Auto-Encoder (VAE).

Data Analysis in Astronomy I

Pipeline for detection of transient objects in optical surveys

Nicolai S Pankov[1]
1 National Research University "Higher School of Economics" (HSE)
nspankov@edu.hse.ru

Identification and following study of optical transients (OTs) associated with cosmic gamma-ray bursts (GRBs) and gravitational wave (GW) events is a relevant research problem of multi-messenger astronomy. The projenitors of OTs are initially localized with space gamma and X-ray telescopes (for GRBs) or ground-based laser interferometers LIGO/Virgo/KAGRA (for GW events). A joint localization area had a range of 10 to 1000 of square degrees in previous cycles of LIGO/Virgo observations. Optical surveys equipped with wide-field cameras allow to cover the entire localization area in the several scans. As the result, the bulky series of survey images are generated. After processing they contain $\sim10^5$ objects among which it is necessary to identify OT of interest. Both the image processing and identification of OT must be performed in real time due to rapid decay of brightness. Software pipelines become relevant to solve these problems. In this paper, a plugin-extensible architecture of the in progress pipeline for detection of OTs is presented. Each currently implemented unit of the pipeline and their algorithms are described. The estimation of the accuracy and performance of the pipeline units are provided. Plans for the further development of the pipeline are discussed.

Cross-matching of large sky surveys and study of astronomical objects apparent in UV-band only

Aleksandra Sergeevna Avdeeva[1]; Sergey Karpov[2]; Oleg Yu. Malkov[1]; Gang Zhao[3]
[1]Institute of Astronomy, RAS
[2]Institute of Physics, Czech Acad. Sci.
[3]Key Laboratory of Optical Astronomy, National Astronomical Observatories, Chinese Academy of Sciences
avdeeva@inasan.ru; karpov.sv@gmail.com; malkov@inasan.ru; gzhao@nao.cas.cn*

Cross-matching of various information sources is a powerful tool that helps to not only enrich and augment the contents of individual ones, but also to discover new and unique objects. In astronomy, cross-matching of catalogues is a standard tool for getting broader information on the objects by combining their data from the surveys performed at different wavelengths, and it allows to solve number of tasks like studying various populations of astronomical objects or investigating the properties of interstellar medium. However, the analysis of objects present in just one catalogue and missing the counterparts in all others is also a promising method that may lead to the discovery of both transients and objects with extreme color characteristics. Here we report on our preliminary search for objects that manifest only in ultraviolet observations by comparing the data from GALEX catalogue with several

other surveys in different wavelength ranges. We describe the selection of representative sky surveys for this task and give the details on the process of their cross-matching and filtering of the results. We also discuss the possible nature of several outstanding objects detected during the analysis, and discuss the potential output of a larger-scale investigation we are planning based on the experience gained during this initial study.

Application of machine learning methods for cross-matching astronomical catalogs

Alexandra A Kulishova[1]
[1]Lomonosov Moscow State University
*sasha_kulishova@mail.ru**

The article focuses on the application of machine learning methods for cross-matching astronomical catalogs. In the work, the generalization of ideas and results of solving the problem posed from similar works. A brief comparison of the methods used is given. There are described the formulation and results of an experiment using machine learning algorithms to solve the problem, as well as a comparative analysis of their results with the neighborhood method. This work can be used to implement the stage of cross-matching of catalogs in modern astronomical systems.

Clustering stellar pairs to detect extended stellar structures

Sergei Sapozhnikov[1]
[1]Institute of Astronomy, RAS
thestriks@gmail.com

Gaia data allows for search for extended stellar structures in phase (coordinates plus velocities) space. We describe a method of using DBSCAN clustering algorithm, which is used to group closely-packed-together data points, to a list of preliminary selected pairs of stars, with parameters expected to be found within stellar streams and comoving groups: loose structures in which stars are not gravitationally bound, but do share motion and evolutionary properties. To test our approach, we construct a model population of background stars, and use pair-constructing and clustering algorithms on it. Results show that transitioning to a list of pairs sharply reveals structures not presented in background model, which then become more apparent targets in coordinates-velocities phase space for DBSCAN algorithm thanks to now increased relative density of the extended stellar structure.

Data Analysis in Astronomy II

Host galaxies of cosmic gamma-ray bursts

Alina Volnova[1]; Alexey S. Pozanenko[1]; Sergey Belkin[2]
[1]Space Research Institute, RAS
[2]National Research University "Higher School of Economics" (HSE)
alinusss@gmail.com; apozanen@iki.rssi.ru; astroboy96@mail.ru

The discovery of gamma-ray burst (GRB) host galaxy back in 1997 brought confirmation of GRBs cosmological origin. Nowadays investigation of the host galaxies often is the only way to estimate the cosvological redshiftof GRB sources. The morphology of host galaxies gives clues to the nature of the environment, where the GRBs were born, and allows estimating physical parameters of the circumburst medium. There are two main methods of host galaxies investigations: spectroscopy and broad-band photometry. The latter method is more common for GRBs hosts, since host galaxies are relatively distant, and their flux is too faint for valuable spectroscopy. We discuss investigations of host galaxies and particularly multicolor photometryof host galaxies. The method is based on the comparison of host galaxies multicolour photometry with the synthetic spectral energy distribution modelled using the theory of the stellar evolution. The best model gives the estimates of main galaxy parameters: redshift, age, mass, morphological type, internal extinction and star formation rate. We also present the results of the modelling of host galaxies from IKI GRB-FuN catalogue and discuss results in the framework of known GRB host galaxies. The increase of the GRB host galaxies statistics including the short duration of GRB will be helpful in the process of selection of target galaxies in search for counterparts of LIGO/Virgo/KAGRA gravitational wave events in next runs.

The diversity of light curves of SNe associated with GRBs

Sergey Belkin[1]; Alexey S. Pozanenko[2]
[1]National Research University "Higher School of Economics" (HSE)
[2]Space Research Institute, RAS
astroboy96@mail.ru; apozanen@iki.rssi.ru

More than 10000 gamma-ray bursts (GRBs) have been detected since discovery. Long-term observations of about 850 GRB afterglow in optic since 1998 have shown that a core-collapse supernova (SN) accompanies about 50 nearby GRB sources. We have collected about two dozen SNe' multicolor light curves associated with GRBs. The sample is based on published data, obtained during observations of GRB-SN cases by ground-based observatories all around the world including our own observations. A description of the procedure for the extraction of the SN's light curve, its analysis, and

phenomenological classification of SNs are presented. We also discuss the current status and problems of investigations of SN associated with GRB.

VALD in astrophysics

Yury V Pakhomov[1]
[1]INASAN
pakhomov@inasan.ru

Vienna Atomic Line Database (VALD) is a most popular among astrophysicists for studying stars, stellar systems, interstellar medium, its chemical composition, evolution and kinematics. The article briefly describes the evolution and modern state of the VALD. The database contains parameters for millions atomic and molecular lines which provide possibility of synthetic spectra calculation. VALD is a active member of international project "Virtual Atomic and Molecular Data Centre" (VAMDC).

Star clusters, planets, asteroids and comets in the light of Big Data

Sergei Vereshchagin[1]; Maria Sizova[1]; Aleksandr Tutukov[1]; Andrei Fionov[2]
[1]INASAN
[2]OCRV (Russian Railway)
svvs@ya.ru; sizova@inasan.ru; atutukov@inasan.ru; fionovs@mail.ru

In clusters, the stellar density is 1000 times higher than the density of field stars and, thus, clusters represent the most promising places for searching for planetary systems (exoplanets and exocomet). It is important to note that objects belonging to clusters are automatically time bound, since the ages of clusters are determined much more reliably than individual stars in the field. The result of such a search is the accumulation of a gigantic amount of information about stars (Gaia), exoplanets (TESS), exocomet. Naturally, the problem arose of comparing data on stars, star clusters, planets and comets. Most of these data represent a stream of information increasing over time: new editions of the Gaia catalogs, the stream of TESS observations. The exponential growth of this data can include the development of methods and tools for analyzing and managing data, learning from the experience of applying these methods in other areas of knowledge. The need to constantly replenish complex data on planetary systems, stars and star clusters has given rise to our work. The proposed structure for working with data allows one to determine the relationships and parameters of the studied systems of the types star - comet, star - exoplanet, star - cluster - planetary system. At the output, data analysis, development and use of infrastructure applicable in various DIDs (data intensive domains) are important in order to obtain and publish new knowledge

Data Analysis in Earth Sciences, Advanced Data Analysis Methods

On the development of a pipeline for processing hydrometeorological data

Evgenii D. Viazilov[1]; Denis Melnikov[1]; Alexander Mikheev[2]
[1]ФГБУ "ВНИИГМИ-МЦД"
[2]RIHMI-WDC
vjaz49@mail.ru; melnikov@meteo.ru; miheev1807@yandex.ru; vjaz@meteo.ru

For the first time in worldwide for hydrometeorology, data processing pipeline is proposed. Approaches to its implementation are defined. The stages and software for such processing are highlighted. The main method of the universal data replenishment mechanism with the results of the pipeline is an integrated database with a wide set of metadata. The creation of information products at various stages of the pipeline is considered. It is proposed to create new services for the pipeline. The main control mechanism for pipeline data processing is considered to be tools of monitoring the state of hardware, software and information resources. This requires a transition from monitoring individual stages to automatic comprehensive monitoring of the data processing pipeline state. Tasks of the administrator of the data processing pipeline are defined. In fact, when using pipeline processing, a transition must be made to fully automatic data processing without human intervention.

Image recognition for large Soil Maps Archive overview: metadata extraction and georeferencing tool development

Nadezda A. Vasilyeva[1]; Artem Vladimirov[1]; Taras Vasiliev[1]
[1]Dokuchaev Soil Science Institute
nadezda.vasilyeva@gmail.com; artem.a.vladimirov@gmail.com; tarasvasiliev44@gmail.com*

During the second half of 20th century Dokuchaev Soil Science Institute has collected soil maps for a half of the Eurasian continent as a result of large national soil surveys which lasted for several decades with the efforts of the former USSR. Such labor-intensive expeditions on countries scale were not repeated since then. The question of future soil dynamics as Earth's fertile layer became crucial with global population growth and causes large part of uncertainty in Earth System Modelling. Most of the present knowledge about soil types is still in form of paper soil maps, representing valuable knowledge about soil cover of the past. Soil type itself is a crucial factor which still cannot be determined remotely but can be updated. Archive soil maps (several thousands of sheets) are an example of data which require digitizing and could profit from application of image recognition techniques. In the current study we present a demo tool for fast extraction of metadata and geo-referencing of paper soil maps using image recognition techniques. Presented software can be used for creating soil maps digital catalog allowing for a quick overview of a large collection.

Evolutionary Approach to Multimodal Clustering on Formal Contexts.

Mikhail Y Bogatyrev[1]; Dmitry Orlov[1]; Sergey Dvoenko[1]; Tatyana Shestaka[1]
[1]Tula State University
okkambo@mail.ru; di-orl@mail.ru; sergedv@yandex.ru; shestaka.tanya@mail.ru

Evolutionary approach to multimodal clustering on multidimensional formal contexts is proposed. Its main advantage is that it allows one to find a well-grounded number of clusters corresponding to the global or close to the global extremum of the function that characterizes the quality of solution of the clus-tering problem. This approach is effective when the proximity measure for a data being clustered is not Euclidean. Formal context is the data model in For-mal Concept Analysis, the area in data analysis where mathematically rigorous methods from lattice theory have been applied for discovering relationships on heterogeneous data. Taking into account the effect of data heterogeneity in cluster analysis can be effectively implemented using multimodal clustering methods. The paper contains main definitions from Formal Concept Analysis, description the principle of evolutionary computation and evolutionary approach to multimodal clustering. Experimental study of proposed approach is performed on the task of phenotyping of disease of myocardial infarction.

Distances Parameterized by Size: Models and Adaptation Techniques

Archil I. Maysuradze[1]; Daria Petrenko[1]
[1]Lomonosov Moscow State University
artchil@mail.ru; daria_petrenko@bk.ru

In many applied domains, the concept of distance is used for initial formulation and subsequent formalization of problems and solution methods. However, for an adequate representation of complex situations, the traditional concept of distance is insufficient, and richer families of models are required. In this paper we propose and investigate theoretically and empirically one of the families - distances parameterized by size. We also introduce the generalized metric axioms as a set of natural requirements in many domains. As examples of applied domains, we can consider transport systems, in which the transportation time depends on the mass of the cargo, or message passing networks, in which the transfer delay depends on the length of the message. The number of combinations of object couples and sizes is huge, so the complete description of all the situations is data intensive.

Then the problem of modelling and approximating the collected dissimilarity tensor is posed and solved in various ways. Several models of distances parameterized by size are proposed in the work. For each of the models, sufficient conditions are found on the parameters (theorems on sufficient conditions) that ensure the fulfillment of all the generalized metric axioms. To adapt each of the models,

we propose a specific method of conditional optimization. The idea of methods is in iterative conditional minimization of the variational upper bound for the stress function.

All the proposed models and methods were implemented and tested on real data on message passing delays between processes in the Lomonosov supercomputer system. Experiments have shown a good quality of approximation for models with a small number of parameters (that is, a high degree of data compression), as well as comparability of losses with unconditional problem statements in which the generalized metric axioms are ignored.

MODELS OF DECISION MAKING WITH LIMITED VOLUME OF PROCESSED INFORMATION

Budzko, M. A. Gorelov, F. I. Ereshko [1]
[1]ВЦ ФИЦ ИУ РАН (1171)
fereshko@yandex.ru

The paper examines a new class of models, the common feature of which is the presence of limited volume of information exchanges between active players, which is typical for subjects associated with agricultural production. It is shown that taking such restrictions into account allows one to significantly expand the class of situations that can be adequately described using the game theory in normal form. Possible ways of formalizing the concept of "amount of information" are discussed returning to the constructions by A.N. Kolmogorov. A new concept of the maximum guaranteed result is discussed. The application of the outlined ideas in the information theory of hierarchical systems is considered.

Information extraction from texts I

Domain-specific Taxonomy Enrichment based on Meta-Embeddings

Mikhail Tikhomirov[1]; Natalia V Loukachevitch[1]
[1]Lomonosov Moscow State University
tikhomirov.mm@gmail.com; louk_nat@mail.ru

In this paper we study the use of meta-embeddings approaches, which combine several source embeddings, for the taxonomy class prediction of new terms. We test the proposed approach in the information-security domain in the task of enriching the Ontology on Natural Sciences and technologies (OENT). We show that autoencoder-based meta-embeddings with triplet loss achieve the best results in the task. The highest results are obtained on combination of in-domain and out-of-domain embeddings.

Extracting Sentiments towards COVID-19 Aspects

Eduard Nugamanov[1]; Natalia V Loukachevitch[1]; Boris V Dobrov[2]
[1]Lomonosov Moscow State University
[2]Research Computing Center of M.V.Lomonosov Moscow State University
ed.nugamanov@gmail.com; louk_nat@mail.ru; dobrov_bv@mail.ru

In this paper, we introduce a specialized Russian dataset and study approaches for aspect-based sentiment analysis of Russian users' comments about the COVID-19. We solve two tasks, namely Relevance Determination (RD), which aims to predict whether a sentence is relevant to an aspect of the pandemic, and Sentiment Classification (SC), which classifies the sentiment expressed towards an aspect in a sentence. We applied and tested various methods of machine learning, including fine-tuning of the pre-trained RuBERT model. The best results in both tasks were obtained by RuBERT model in the Natural Language Inference (NLI) formulation.

Cross-lingual plagiarism detection method

Denis Zubarev[1]; Ilya Tikhomirov[2]; Ilya V Sochenkov[3]
[1]PFUR
[2]Russian Science Foundation
[3]Federal Research Center "Computer Science and Control" of Russian Academy of Sciences
zubarev@isa.ru; tia@rscf.ru; sochenkov@isa.ru

In this paper, we describe a method for cross-lingual plagiarism detection for a distant language pair (Russian-English). All documents in a reference collection are split into fragments of fixed size. These fragments are indexed in a special inverted index, which maps words to a bit array. Each bit in the bit array shows whether a Ith sentence contains this word. This index is used for the retrieval of candidate fragments. We employ bit arrays stored in the index for assessing similarity of query and

candidate sentences by lexis. Before doing retrieval, top keywords of a query document are mapped from one language to other with the help of cross-lingual word embeddings. We also train a language-agnostic sentence encoder that helps in comparing sentence pairs that have few or no lexis in common. The combined similarity score of sentence pairs is used by a text alignment algorithm, which tries to find blocks of contiguous and similar sentence pairs. We introduce a dataset for evaluation of this task - automatically translated Paraplag (monolingual dataset for plagiarism detection). The proposed method shows good performance on our dataset in terms of F1. We also evaluate the method on another publicly available dataset, on which our method outperforms previously reported results.

An approach to processing news text messages based on markeme analysis

Alexander V. Sychev[1]
[1]Voronezh State University
sav@sc.vsu.ru

The complexity problem of automatic message filtering retrieved from online media platforms and social networks is discussed. An approach to the text of news messages processing based on the markeme analysis is suggested in the paper. Markemes identification is based on calculating the Index of Textual Markedness (InTeM). Markemes are words most important for a particular text and occur with the frequency, which is higher than that of the words of the same length.Application of the approach under proposal to the news messages clustering problem is studied in the paper. Preliminary results of the proposed approach study, which are related to the news messages classification and clustering problems, are presented and discussed.

Information extraction from texts II

Methods for automatic argumentation structure prediction

Ilya Dimov[1]; Boris V Dobrov[2]
[1]Lomonosov Moscow State University
[2]Research Computing Center of M.V.Lomonosov Moscow State University
iliyadimov@icloud.com; dobrov_bv@mail.ru

Argumentation mining is a natural language understanding task consisting of several subtasks: relevance detection, stance classification, argument quality assessment and fact checking. In this work we propose several architectures for the analysis of argumentative texts based on BERT. We also show, that models, which jointly learn argumentation mining subtasks outperform pipelines of models trained on a single tasks. Additionally we explore transfer learning approach based on pretraining for the natural language inference task, which achieves highest score on tasks of argumentation mining among the models trained on english corpora.

Improving Neural Abstractive Summarization with Reliable Sentence Sampling

Daniil Chernyshev[1]; Boris V Dobrov[2]
[1]M.V.Lomonosov Moscow State University
[2]Research Computing Center of M.V.Lomonosov Moscow State University
chdanorbis@yandex.ru; dobrov_bv@mail.ru

State-of-the-art abstractive summarization models are able to produce summaries for various types of sources with quality comparable to human written texts. However, despite the fluency, the generated summaries are often erroneous due to factual inconsistencies caused by neural hallucinations. In this work, we study possible ways of reducing the hallucination rate during abstractive summarization. We compare three different techniques aimed at improving the correctness of the training procedure: control tokens, truncated loss, and dataset cleaning. To control hallucination rate outside of the training, we propose an improved algorithm for summary sampling - reliable sentence sampling. The algorithm utilizes fact precision metrics to sample the most reliable sentences for an abstractive summary. By conducting the human evaluation, we demonstrate the algorithm's efficiency in preserving summary factual consistency.

Search query extension semantics

Olga Ataeva[1]; Vladimir Serebryakov [1]; Natalia P Tuchkova[2]
[1] Dorodnicyn Computing Center FRC CSC of RAS
[2]FRC CSC RAS
oli@ultimeta.ru; serebr@ultimeta.ru; natalia_tuchkova@mail.ru*

The problems of extracting the most complete information from the semantic library by accounting for related documents are considered. Expert knowledge encrypted in the subject area can

be made available when the user obtains additional information from linked documents. A feature of the approach in the application of the algorithm of a shallow neural network to expand the search query and mathematical subject areas, where expert knowledge is available with a significant scientific knowledge of users. The solution to this problem can be achieved by means of semantic analysis in the knowledge space using machine learning algorithms. The paper investigates the construction of a vector representation of documents based on paragraphs in relation to the data array of the digital semantic library LibMeta. Each piece of text is labelled. Both the whole document and its separate parts can be marked.

Since the problem of enriching user queries with synonyms was solved, when building a search model, together with word2vec algorithms, an approach of "indexing first, then training" was used to cover more information and give more accurate results.

DEVELOPMENT OF METHODS FOR EXTRACTING INFORMATION FROM PHARMACY LINE USING CONDITIONAL RANDOM FIELDS

Evgenia Mitrokhina [1]; Alexey I. Molodchenkov [1,2,3]; Artem A. Nikolaev [3]
[1]Moscow Institute of Physics and Technology, Dolgoprudny
[2]Federal Research Center "Informatics and Management" of the Russian Academy of Sciences, Moscow
[3] Peoples' Friendship University of Russia, Moscow
mitrohina.ea@phystech.edu; aim@tesyan.ru; nicepeopleproject@gmail.com*

The paper considers the solution to the problem of extracting information from short lines of pharmacological orientation in Russian language. As an example, pharmacy lines are used, from which you need to extract the full name of the drug, manufacturer, form of issue, dosage, number of pieces in a package and some other parameters. To extract this information, a conditional random field (CRF) algorithm was used. There was also created a method for preliminary standardization of the strings to bring string tokens to a single form. More than seven thousand pharmacy lines were marked for the experiments and 2 CRF models were trained - with and without preliminary standardization of the lines. For the model with standardization, the following results were obtained: accuracy for different data sets is 0.95 (on the validation set) and 0.89 (on the test set). For the model without standardization, the accuracy is 0.95 (on the validation set) and 0.87 (on the test set).

Scientific and Educational Texts Analysis

A system for information extraction from scientific texts in Russian

Elena Bruches[1]; Anastasia Mezentseva[1]; Tatiana V Batura[1]
[1]Novosibirsk State University
*bruches@bk.ru; anastasiamez@mail.ru; tatiana.v.batura@gmail.com**

In this paper, we present a system for information extraction from scientific texts in the Russian language. The system performs several tasks in an end-to-end manner: term recognition, extraction of relations between terms, and term linking with entities from the knowledge base. These tasks are extremely important for information retrieval, recommendation systems, and classification. The advantage of the implemented methods is that the system does not require a large amount of labeled data, which saves time and effort for data labeling and therefore can be applied in low- and mid-resource settings. The source code is publicly available and can be used for different research purposes.

Development of Lexico-Syntactic Ontology Design Patterns for Information Extraction of Scientific Data

Kristina Ovchinnikova[1]; I. Kononenko[2]; Elena A. Sidorova[2]
[1] Novosibirsk State University
[2]A.P. Ershov Institute of Informatics Systems SB RAS
*k.ovchinnikova2@g.nsu.ru; irina_k@cn.ru; lena@iis.nsk.su**

The work considers an approach of information extraction based on lexico-syntactic patterns (LSPs). LSPs are built on the basis of knowledge about the sci-entific subject domain presented in the ontology and the corpus of scientific pub-lications of different areas of knowledge. Two key tasks must be solved with the help of the LSPs: extracting the names of objects and constructing objects in ac-cordance with the structure of the ontology classes. In line with these tasks, ter-minological and informational LSPs are differentiated. Terminological patterns ensure the extraction of object names and properties based on indicators - marker words and phrases. Information patterns provide identification of ontology ob-jects based on key attributes, description of actant structure for predicates ex-pressing attributive relations and relations between ontology objects, as well as matching language constructions to values of attributes of ontology objects and their relations. Research is conducted on the basis of a corpus of scientific publi-cations, which includes 100 articles from various fields of knowledge. The ways of expressing information about research methods as the central concept of the ontology of scientific activity are investigated

Dask based efficient clustering of educational texts

Fail Gafarov[1]; Dmitriy Minullin[1]; Viliuza Gafarova[2]
[1]Kazan Federal University
[2]TR AS Institute of Applied Semiotics
fgafarov@yandex.ru; minullin.dima@mail.ru; 79046639045@yandex.ru

Document clustering process is a long running and computationally demanding process. The need for systems that allow fast document clustering is especially relevant for processing large volumes of text data (Big Data). In this work we present a distributed text clustering framework based on Dask open source library for parallel and distributed computing. The Dask-based processing system developed in this work allows to execute all necessary operations related to the clustering of text documents in a parallel mode. We realized parallel agglomerative clustering algorithm of cosine similarity matrices computed from term frequency-inverse document frequency (TF-IDF) feature matrices of input texts. The system had been applied to intellectual analysis of educational data accumulated in the system "Electronic education of the Tatarstan Republic" from 2015 to 2020. Specially, by using developed system we clustered the text documents describing lesson planning, and also performed a comparative analysis of the average marks of students, whose training was carried out according to lesson planning belonging to different clusters.

Cognitive processes in generating and restoring elliptical sentences

Xenia A. Naidenova[1]
[1]S. M. Kirov Military Medical Academy, Saint-Petersburg
ksennaidd@gmail.com

A new cognitive approach to resolving ellipses in geometry texts is advanced. This approach is evolving in an automated system for solving school geometry tasks expressed in natural Russian language. A classification of ellipses occur-ring in geometry texts is given and the rules of converting the complete sentences to their elliptical variant are formulated. The cognitive schemes are introduced as the syntactic synonyms of sentences describing planimetric configurations. The role of cognitive schemes in understanding sentences is considered. They are represented by the drawing and NL-texts generated on the principle of combining the noun, verb, and prepositional phrases corresponding with both the fragments of schemes and the expressions in real geometric texts. The cognitive schemes of geometric configurations allow to facilitate the process of syntactical and semantical parsing the tasks' text, to resolve ellipses, to visualize the task condition and to reveal hidden geometric relationships not explicitly expressed in the text of task.

Methods and Tools for Literary Texts Analysis

Analyzing the cultural universals of the Folklore of Peoples of Siberia and the Far East

Anna A. Grinevich[1]; Alexey Sery[2]
[1]Institute of Philology SB RAS
[2]A.P. Ershov Institute of Informatics Systems, Siberian Branch of the Russian Academy of Sciences
annazor@mail.ru; alexey.seryj@iis.nsk.su*

The article describes an ontological approach to presentation of folklore of the Siberian peoples from the point of view of cultural universals, which are both widespread concepts that form the linguistic picture of the world and the methods of preserving oral traditional culture, as well as the poetics and form of folklore phenomena. The proposed approach is to formalize the subject domain of cultural universals using one or more ontologies and to build an information system providing tools for research and analysis on the basis of these ontologies. The paper discusses the designed domain ontologies and development of the information system. An attention is paid to how the system provides access to its resources, which are folklore works of various kinds, and how it keeps ontologies.

Using a decision tree to identify non-uniform fragments in a text

Alexander Rogov[1]; Kirill Kulakov[1]; Nikolai Moskin[1]; Roman Abramov[1]
[1]Petrozavodsk State University
rogov@petrsu.ru; kulakov@cs.karelia.ru; moskin@petrsu.ru; monset008@gmail.com

This article discusses the problem of searching for non-uniform text fragments. It can be several paragraphs or individual sentences that differ significantly from the rest of the text in terms of a set of characteristics. The problem of finding non-uniform fragments and their interpretation arises in the study of the pre-revolutionary magazines "Time" (1861-1863), "Epoch" (1864-1865) and the weekly "Citizen" (1873-1874). As you know, F. M. Dostoevsky was their editor, as a result of which he could make his own edits to the texts of articles by other authors. In our research the texts were divided into separate parts, for each of which the frequency of n-grams (encoded sequences of parts of speech) was determined. Further, the analysis was carried out using decision trees that classified texts by author. In particular, the texts of F. M. Dostoevsky and V. P. Meshchersky were subjected to a similar analysis.

Technological features of cross-language migration from PHP to Python of software products working with intensive data

Vladimir B Barakhnin[1]; Olga Kozhemyakina[2]; Artem Revun[1]; Natalia Shashok[3]
[1]Federal Research Center for Information and Computational Technologies
[2]ICT SB RAS
[3]NSU
barakhnin@ngs.ru; olgakozhemyakina@mail.ru; FarlLionDirk@yandex.ru; n.shashok@g.nsu.ru

This article describes the methods of cross-language migration from the PHP programming language to the Python programming language on the exam-ple of modernization of the search module of the system for complex analy-sis of poetic texts. The choice of language was determined by the need of in-tegration of the module in the system, and by the fact that Python has a large number of open libraries that make it possible to work with machine learning technologies, what is an advantage in the intensive data processing. The main problems of cross-language migration are considered and the methods to solve these problems are proposed. The proposed solutions con-cern the data security, the separation of module representation from the data model, the work with database via declarative mapping. In the process of cross-language migration, the search module was rewritten and all problems were solved, what allowed to integrate it into the information system for complex analysis of poetic texts.

Alphabetical Index