10th Russian Summer School in Information Retrieval (RuSSIR 2016)

Alexander Kotov
Wayne State University, MI, USA
kotov@wayne.edu

Elena Treshcheva
Saratov State University, Russia
etres@sgu.ru

Leonid Bessonov
Saratov State University, Russia
lexx@sgu.ru

Dmitry I. Ignatov
Higher School of Economics, Russia
dignatov@hse.ru

Yana Volkovich
AppNexus, NY, USA
yvolkovich@appnexus.com

Maria Eskevich
Radboud University, The Netherlands
m.eskevich@let.ru.nl

Pavel Braslavski
Ural Federal University, Russia
pbras@yandex.ru

1 Introduction

The 10th Russian Summer School in Information Retrieval (RuSSIR 2016) was held on August 22–26 in Saratov, Russia[1]. The school was co-organized by Saratov National Research State University[2] and the Russian Information Retrieval Evaluation Seminar (ROMIP)[3].

The RuSSIR school series started in 2007 and has evolved into a renowned academic event with extensive international participation [1, 2, 3]. Previously, RuSSIR has taken place in Yekaterinburg, Taganrog, Petrozavodsk, Voronezh, St. Petersburg, Yaroslavl, Kazan, and Nizhny Novgorod. Over the years, many RuSSIR courses have been taught by world-renowned researchers in Information Retrieval (IR) and related areas.

Saratov is located in the European part of Russia. Situated on the Volga river, it is known for its 3-km-long beautiful bridge connecting Saratov and Engels. Saratov has many cultural and historical sites. The Volga embankment and Kirov avenue are the two main streets for a pleasant walk along the river and for exploring the city.

Saratov State University (SSU) has the status of a National Research University, with long-standing excellent educational programs in Mathematics and Computer Science. The SSU pro-

[1] http://romip.ru/russir2016/
[2] http://www.sgu.ru/en/
[3] http://romip.ru/en/

gramming teams have won silver medals in the 2002 and 2003 ACM ICPC World Competitions and also the European programming championship in 2002. In 2006, SSU students won first prizes in both the European Programming Championship and the ACM Contest World Finals.

Saratov is the birthplace of the prominent 19th century Russian philosopher and writer Nikolay Chernyshevsky. Saratov State University was named after him. The quote "*In the name of their beloved science... they were killing an enormous number of frogs*" from Chernyshevsky's most famous novel "*What is to be done?*" was printed on RuSSIR 2016 T-shirts.

Along with courses on traditional IR topics, such as click models, personalization, clustering and recommender systems, the RuSSIR 2016 program featured a special track of courses focusing on semantic search, construction of semantic repositories (entity recognition, information extraction, knowledge etc.) and their utilization in IR. The program led to many fruitful discussions among the participants coming from different areas and allowed students to expand their knowledge across multiple disciplines. The school program included one keynote lecture, nine short and regular courses running in two parallel sessions, one sponsor talk, and the RuSSIR 2016 Young Scientist Conference.

The school welcomed 78 participants, who were selected based on their applications. The majority of students were from Russia (Volgograd, Vologda, Dolgoprudny, Ekaterinburg, Izhevsk, Kazan, Krasnoyarsk, Moscow, Perm, Saint Petersburg, Saratov, Tyumen, Chelyabinsk, Yaroslavl). There were also 14 students from other countries: Czech Republic, France, Greece, India, Iran, Ireland, Latvia, Portugal, Ukraine. The audience of RuSSIR consisted of undergraduate, graduate and doctoral students, as well as young faculty and industry professionals. The total number of participants including students, sponsor representatives, lecturers and organizers was 121.

School participation was free of charge due to support from the school's sponsors. 20 accommodation grants were awarded to Russian and international participants. Additionally, five European-based students as well as three school lecturers from Europe and the United States received travel support from the European Science Foundation (ESF)[4] through ELIAS network[5].

2 Courses

The program of RuSSIR 2016 was compiled based on invited courses along with course proposals reviewed by three Program Committee (PC) members and accepted by the PC Chair. In total, six course proposals were submitted, four of which were selected for inclusion in the school program. The special track of RuSSIR 2016 included five invited courses, each of which consisted of two 90-minute lectures. The four selected courses consisted of four or five 90-minute lectures taught throughout the school week. All courses ran in two parallel sessions. Short descriptions of each course are provided below.

Computational Social Science: Theories, Methods and Data – Ingmar Weber, Qatar Computing Research Institute, Qatar

Dr. Weber's course was designed to bridge the gap between Sociology and Computer Science. On one hand, he provided an introduction to social theories and models helping students to better

[4] http://www.esf.org/

[5] http://www.elias-network.eu/

understand the processes generating social data. On the other hand, he provided an overview of statistical and computational methods which are useful for addressing social science research questions, with a particular focus on the Web observational data. The course covered an extensive set of techniques to not only answer "how", but also "why" questions. In particular, it provided an overview of important sociological theories, how to derive verifiable hypotheses from them and methods to use for causal inference from observational data to go beyond mere correlations. Course learning objectives were reinforced by a set of short practical examples provided in the form of IPython notebooks as well as by a small competition, in which students were invited to submit their own research proposals. This course is an extended version of a tutorial given by Dr. Weber and his colleagues at the 25th International World Wide Web Conference (WWW 2016).

Knowledge Graph Entity Representation and Retrieval – Alexander Kotov, Wayne State University, USA

Recent studies indicate that more than 75% of queries issued to Web search engines aim at finding information about entities, which could be material objects or concepts that exist in the real world or fiction (e.g. people, organizations, locations, products, etc.). Most common information needs underlying this type of queries include finding a certain entity (e.g. "Einstein relativity theory"), a particular attribute or property of an entity (e.g. "Who founded Intel?") or a list of entities satisfying a certain criteria (e.g. "Formula 1 drivers that won the Monaco Grand Prix"). These information needs can be efficiently addressed by presenting structured information about the target entity or a list of entities retrieved for these queries from a knowledge graph, either directly as search results or in addition to the ranked list of documents. The course taught by Dr. Kotov provided a summary of the latest research in knowledge graph entity representation methods and retrieval models. The first part of this course introduced different methods for entity representation: from multi-fielded documents with flat and hierarchical structure to latent dimensional representations based on tensor factorization. In the second part of this course, Dr. Kotov provided an overview of recent developments in entity retrieval models, including Mixture of Language Models (MLM), Probabilistic Retrieval Model for Semi-structured Data (PRMS), Fielded Sequential Dependence Model (FSDM) and its parametric extension (PFSDM) as well as learning-to-rank methods.

Entity Linking – Krisztian Balog, University of Stavanger, Norway

In his course, Dr. Balog focused on the problem of entity linking, which refers to identifying and disambiguating entity mentions in text and linking them to their corresponding entries in a reference knowledge base. Entity linking has proven to be the key step towards understanding the meaning of text. Entity annotations allow readers of a document to acquire contextual or background information with a single click. They can also be used in downstream processing to improve retrieval performance or to facilitate better user interaction with documents or search results, e.g., by providing easy access to related entities. The course by Dr. Balog consisted of two lectures: in the first lecture, he introduced the theory and methods for entity linking, while in the second lecture, he provided practical guidance with hands-on examples and exercises.

Ontological Information Extraction – Fabian Suchanek, Telecom ParisTech University, France

The course taught by Dr. Suchanek provided a detailed overview of information extraction

techniques – techniques that are used in the process of deriving structured information from text. The course consisted of two consecutive lectures and focused on factual and semantic information extraction. In particular, the course covered such topics as knowledge representation on the Semantic Web, named entity recognition, entity disambiguation, named entity annotation as well as instance and fact extraction. Dr. Suchanek further provided examples of commercial applications for Information Extraction, such as Google's Knowledge Graph and IBM's Watson question answering system, and of academic projects, such as YAGO, DBpedia, and NELL.

Domain Specific Semantic Search – Mihai Lupu, Vienna University of Technology, Austria

Dr. Lupu's course started with an introduction to domain-specific search and relevant methods including adapting vector space weighting and applying specialized vocabularies. Next, Dr. Lupu focused on addressing the specific challenges arising in medical and patent IR. In the first case, the quality and trustworthiness of documents together with their readability level, i.e. is information understandable by a layperson or only by a medical professional, is of extreme importance. In the second case, professional patent searchers are faced with the task of finding all patents related to a query (i.e. high recall task) despite the fact that patent documents are often not written to be easily found. Horizontal to these two domains is the issue of credibility in IR, which was also discussed in detail. In the last part of the course, Dr. Lupu provided an overview of IR evaluation, evaluation campaigns and their results for medical and patent IR tracks.

Social Personalization and Recommender Systems – Shlomo Berkovsky, The Commonwealth Scientific and Industrial Research Organization (CSIRO), Australia

The course by Dr. Berkovsky addressed the problem of information abundance, which often prevents users from discovering desired information or complicates making informed decisions. In particular, it highlighted the pressing need for intelligent personalized applications that simplify information access and discovery by providing adaptive services based on the preferences and needs of their users and focused on recommender systems, one type of personalized application that has recently become tremendously popular in research and industry. Such systems provide personalized recommendations to users about information and products they may be interested in examining or purchasing. This is often achieved by exploiting social methods, which amalgamate past experiences of other users in order to identify the most valuable information and products. Extensive research into recommender systems over the last decade has yielded a wide variety of techniques, which have been published at a range of reputable venues and subsequently adopted by numerous Web-sites and services. This course provided a broad overview of algorithms and techniques for recommender systems and practically deployed Web and mobile applications of personalized technologies.

Knowledge Base Population – Heng Ji, Rensselaer Polytechnic Institute, USA

In her course, Dr. Ji introduced state-of-the-art Information Extraction (IE) and Knowledge Base Population (KBP) methods. In the first lecture, Dr. Ji focused on quality issues for IE and KBP, reviewed the most successful recently proposed methods and discussed the remaining problems. The second lecture covered portability issues, i.e. how to build a new IE/KBP system for a new language, domain, or genre within a short time and at a low cost. A brand new "Liberal" Information Extraction (IE) paradigm was introduced to combine the merits of traditional IE

(high quality and fine granularity) and Open IE (high scalability). Liberal IE aims were to discover schemas and extract facts from any input corpus without any annotated training data or predefined schema.

Click Models for Web Search – Ilya Markov, University of Amsterdam, Netherlands; Aleksandr Chuklin, Google, Switzerland & University of Amsterdam, Netherlands; Maarten de Rijke, University of Amsterdam, Netherlands

This course focused on click models and probabilistic models of interaction between search engines and users, topics that have been extensively studied by the information retrieval community in recent years. We now have a handful of click models, parameter estimation methods, evaluation principles and applications of click models that form the building blocks of ongoing research efforts in this area. Click models also appear in many other areas of IR, such as ranking, evaluation, user simulation, etc. This course covered a wide range of topics, from basic and advanced click models to click model estimation, evaluation and applications of click models, and is based on a recent book on this topic written by the presenters[6]. Most topics were augmented with live demos, where the participants could try the presented material in practice. Furthermore, the course featured two practical sessions, where participants had a chance to implement basic and advanced click models using open-source tools and publicly available datasets with click logs.

Model- and Experiment-driven Recommendations for Haunting Issues in Clustering – Boris Mirkin, Higher School of Economics, Russia & Birkbeck, University of London, UK

In his course, Dr. Mirkin demonstrated that, although clustering is a well-studied problem in the area of data and text analytics, it still has a number of unsolved issues and concerns. In particular, the course provided answers to the following questions, among others: (i) how to find out if there is any clustering structure in a given dataset at all? (ii) if there is, how many clusters are there? (iii) what object-to-object similarity measure one can choose? (iv) which features are useful for clustering and which ones are not? (v) can a mixed, categorical and numerical, feature space be used for clustering? (vi) how can one reconcile clustering solutions that are different from each other? Dr. Mirkin discussed both well-known methods and novel developments in the theory of clustering, including the k-means algorithm, Ward divisive clustering, one-cluster clustering, consensus clustering, spectral clustering, and network community detection. In the first lecture, Dr. Mirkin focused on object-to-feature conversion and in the second lecture, he discussed object-to-object (dis)similarity measures.

Sponsor's Lecture

Georgii Ivanov from *Mail.Ru Group*, one of the RuSSIR 2016 sponsoring organizations, gave a talk on how to improve spelling correction using team draft interleaving data.

3 Young Scientist Conference

For the 10th time, the RuSSIR Young Scientist Conference was part of the school program. The conference helped to create a dialog between school participants and the lecturers. The conference

[6] `http://clickmodels.weebly.com/the-book.html`

call for papers invited young researchers from different areas, such as mathematics and computer science to submit their ideas and research results for various areas of information retrieval. There were two types of submissions: full papers that underwent a thorough reviewing process and short poster notes. Out of 8 submitted full papers, 4 were accepted for publication in the school proceedings:

- Polina Panicheva, Olga Bogolyubova, and Yanina Ledovaya "Revealing Interpretable Content Correlates of the Dark Triad Personality Traits"
- Dmitry Frolov "Using Annotated Suffix Trees for Fuzzy Full Text Search"
- Tatiana Litvinova, Olga Litvinova, Pavel Seredin, Ekaterina Ryzhkova "Deception Bank: A Russian Corpus for Automated Deception Detection in Text"
- Marek Modrý and Michal Ferov "Enhancing LambdaMART using oblivious trees"

The conference ran over two consecutive evenings and consisted of two poster sessions, during which the participants had an opportunity to discuss and exchange their research results and ideas. In total 64 posters were displayed. As in previous years, the Young Scientist Conference was one of the main highlights of the school.

4 Social Program

On the first evening of the school, the RuSSIR Welcome Reception was held in the "Polyglot" cafe. The reception helped school participants to get to know each other. Additional activities on the second and the third evenings, included guided walking tours around Saratov and along the Volga embankment. As a long-standing RuSSIR tradition, a sports event (this year it was beach volleyball) was one of the highlights of the last day of the school. The closing party was held right after the sport competition in an open-air restaurant located by the Volga river. The school's extracurricular activities helped the participants to not only make valuable scientific and personal connections, but to also explore their musical talents including singing, dancing, and playing musical instruments.

5 School Proceedings

For the third time the RuSSIR proceedings are scheduled to be published in the Springer Communications in Computer and Information Science (CCIS) series.[7] The volume will feature two sections: lecture notes and four selected revised papers from the associated Young Scientist Conference. The previous proceedings were published also by Springer CCIS vol. 505 [4] and vol. 573 [5].

[7] http://www.springer.com/series/7899

6 Conclusions

The 10th Russian Summer School in Information Retrieval was a very successful event in many aspects. The school brought together participants with diverse backgrounds from Russia and abroad and facilitated cross-disciplinary exchange of experience and ideas. The RuSSIR students had a unique opportunity to learn new material that is not usually present in university curricula and receive feedback from their peers and lecturers during the poster sessions and informal communications. The event contributed to supporting a lively IR community in Russia and establishing ties with international colleagues. The organizers received very positive evaluation from attendees on various aspects of the school. In 2017, RuSSIR is scheduled to return to Ural Federal University, Yekaterinburg, where the school series started ten years ago.

7 Acknowledgments

We thank all Local Organizing Committee members, Leonid Kossovich, Irina Kirillova, Natalya Stepanova, and Pavel Dmitriev, for their hard work making the school possible, as well as all RuSSIR 2016 Program Committee members for their time and effort to ensure high quality program for RuSSIR 2016, and, in particular, all the lecturers and students who came to Saratov and made the school such a success. We also thank student volunteers who contributed to school organization on site. Our special gratitude goes to Maxim Gubin, who was responsible for legal and financial matters.

We appreciate generous financial support from our sponsors: Yandex[8] and Mail.Ru[9] (golden level), ExactPro Systems[10] and JetBrains[11] (bronze level), Gigant Computer Systems[12] (partner). We also thank the ELIAS network[13] of the European Science Foundation for providing travel grants for RuSSIR lecturers and students and Springer representatives, namely Alfred Hofmann and Aliaksandr Birukou, for their cooperation. Dr. Weber's travel expenses were partly covered by the ACM Distinguished Speaker Program[14].

References

[1] Pavel Braslavski, Nikita Zhiltsov, Stefan M. Rüger, Yana Volkovich: 7th Russian Summer School in Information Retrieval (RuSSIR 2013). SIGIR Forum 47(2): 96-100 (2013)

[2] Pavel Braslavski, Nikolay Karpov, Marcel Worring, Yana Volkovich, Dmitry I. Ignatov: 8th Russian Summer School in Information Retrieval (RuSSIR 2014). SIGIR Forum 48(2): 105-110 (2014)

[8] http://yandex.com
[9] http://go.mail.ru/
[10] http://www.exactprosystems.com/
[11] http://www.jetbrains.com/
[12] http://www.gigant.pro/
[13] http://www.elias-network.eu/
[14] http://dsp.acm.org/

[3] Pavel Braslavski, Ilya Markov, Panos M. Pardalos, Yana Volkovich, Sergei Koltsov, Olessia Koltsova, Dmitry I. Ignatov: 9th Russian Summer School in Information Retrieval (RuSSIR 2015). SIGIR Forum 49(2): 72-79 (2015)

[4] Pavel Braslavski, Nikolay Karpov, Marcel Worring, Yana Volkovich, Dmitry I. Ignatov: Information Retrieval – 8th Russian Summer School, RuSSIR 2014, Nizhniy Novgorod, Russia, August 18-22, 2014, Revised Selected Papers. Communications in Computer and Information Science 505, Springer 2015

[5] Pavel Braslavski, Ilya Markov, Panos Pardalos, Yana Volkovich, Dmitry I. Ignatov, Sergei Koltsov, Olessia Koltsova: Information Retrieval – 9th Russian Summer School, RuSSIR 2015, Saint Petersburg, Russia, August 24-28, 2015, Revised Selected Papers. Communications in Computer and Information Science 573, Springer 2015