8th Russian Summer School in Information Retrieval (RuSSIR 2014)

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

Nikolay Karpov
National Research University
Higher School of Economics, Russia
nkarpov@hse.ru

Marcel Worring
University of Amsterdam, The Netherlands
m.worring@uva.nl

Yana Volkovich
Cornell Tech, US
yv34@cornell.edu

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

1 Introduction

The 8th Russian Summer School in Information Retrieval (RuSSIR 2014) was held on August 18-22, 2014 in Nizhniy Novgorod, Russia.[1] The school was co-organized by the National Research University Higher School of Economics[2] and the Russian Information Retrieval Evaluation Seminar (ROMIP).[3]

The RuSSIR school series started in 2007 and has developed into a renowned academic event with solid international participation. Previously, RuSSIR took place in Yekaterinburg, Taganrog, Petrozavodsk, Voronezh, Saint Petersburg, Yaroslavl, and Kazan. RuSSIR courses were taught by many prominent international researchers in IR and related areas.

Nizhny Novgorod is a beautiful historical city located about 400 km east of Moscow. The city lies on the confluence of the Volga and the Oka rivers. The population of the city is about 1.3 million people. The 16th century Kremlin is the central historic citadel of the city. Nizhny Novgorod is the birthplace of a prominent Russian 20th century writer Maxim Gorky, whose quote "*Science is the sublimest madness of the humankind*" was printed on school t-shirts. Rostislav Alexeev (1916–1980), Russian inventor and designer of hydrofoils and ekranoplans, also lived and worked in Nizhny Novgorod. A picture of "*Raketa*" hydrofoil was printed on school bags.

[1] http://romip.ru/russir2014/
[2] http://www.hse.ru/en/
[3] http://romip.ru/en/

The 2014 RuSSIR programme featured a track of courses focusing on visualization for information retrieval along with other topics, related to information retrieval. The program led to fruitful discussions among participants coming from different domains and allowed students to learn cross-disciplinary competencies. The school programme consisted of two invited lectures, six courses running in two parallel sessions, two sponsor talks, and the RuSSIR 2014 Young Scientist Conference.

The school welcomed 91 participants selected based on their applications. The majority of students came from Russia, but there were also 12 students from European Union and 8 from the rest of the world. The RuSSIR audience comprised of undergraduate, graduate, and doctoral students, as well as young academic faculty and industrial developers. The total number of participants including students, sponsor representatives, lecturers and organisers was 119.

School participation was free of charge thanks to the sponsorship support. In addition, 20 accommodation grants were awarded to Russian participants and 9 European-based students received travel support from European Science Foundation (ESF)[4] through ELIAS network[5]. Travel expenses of three school teachers from Europe were also funded through ELIAS/ESF grant.

2 Courses

The RuSSIR programme was compiled based on reviewing of submitted course proposals by the programme committee. Each course proposal was reviewed by at least six PC members. In total, seventeen course proposals were submitted, six of which were selected for the school programme. Additionally, there were two invited lectures. Each of the six courses consisted of five 90-minute lectures taught in five subsequent days. The invited lectures ran as plenary sessions, the other six courses ran in two parallel sessions.

Seeking Simplicity in Search User Interfaces – Marti Hearst, UC Berkeley, USA

There is a number of difficulties for a new user interface especially in information-intensive tasks like search, consensus finding, or knowledge building to break through and become successful. On the one hand the most complex interfaces might end up unused, but on the other hand successful solutions often lie in a previously unexplored part of the interface design space. In this talk Marti Hearst provided several examples of such successful solutions and discussed possible ideas for future research directions. Dr. Hearst gave her presentation off-site and it was streamed over the Internet. Such online course was a novelty for RuSSIR.

Multimedia Analysis and Multimedia Visualization – Marcel Worring, University of Amsterdam, The Netherlands

There is a semantic gap between retrieval of textual information and multimedia content such as images or videos; it is difficult to get access to their semantics and properly employ their context such as tags, geolocation, and other metadata. Dr. Worring and his colleagues proposed multimedia analytics solutions, which allow to bring human experts and machine algorithms together in a synergetic manner, where the machine performs computationally intensive processing and the expert deals with difficult decisions. Dr. Worring considered methods which learn the semantics of

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

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

images from the interaction with the users in combination with advanced visualizations. Finally, he presented methods for interactive summarizing and exploring of large image collections, e.g. multimedia pivot tables.

Large Scale Information Retrieval – Katja Hofmann, Microsoft Research, Cambridge, UK

Online experimentation focuses on insights that can be gained from user interactions with information retrieval (IR) systems. This course discussed established and recently developed online experimentation techniques, including A/B testing, bandit approaches, interleaving, counterfactual analysis, and online learning to rank. Students were offered to learn about the advantages and limitations of these methods, and how to design online experiments. Finally, a practical component walked students through the process of implementing their own experiments and analyzing obtained results. This practical part featured experiments with freely available Python tools for scientific computing, text parsing, and learning to rank along with relevant data samples.

Web as a Corpus: Going Beyond the n-gram – Preslav Nakov, Qatar Computing Research Institute, Doha, Qatar

The tutorial offered an introduction to computational linguistics and natural language processing, with focus on corpus-based statistical approaches that use the Web as a corpus. Special emphasis was put on Web-based approaches that go beyond the n-gram. Dr. Nakov discussed the reliability and stability of page hits when it is used as a proxy for n-gram frequencies. Various applications of these ideas were explored, from purely linguistic such as resolving syntactic and semantic ambiguities, to more practical but auxiliary tasks such as semantic relation extraction, ontology learning, and search engine query segmentation, to end-applications such as machine translation.

Visualization & Data Mining for High Dimensional Data – Alfred Inselberg, Tel Aviv University, Israel and Pei Ling Lai, Southern Taiwan University of Science and Technology, Tainan, Taiwan

During this course Alfred Inselberg demonstrated an approach to visualization and data mining with parallel coordinates that breaches the perceptual barrier imposed by our 3-dimensional habitation and enables the visualization of multidimensional problems. The goal of such visualization is to systematically incorporate human outstanding ability to pattern-recognition into the problem-solving process. The geometric intuition leads from the patterns initially discovered visually in 3D to the mathematical proofs for the higher dimensions. These patterns are suitable for visual analysis of Big Data and Information Retrieval in a query-result manner; they persist in the presence of errors, and have numerous applications, e.g. collision avoidance for air traffic control, non-linear interactive models of complex systems like a countrys economy and decision support for intensive care units, as well as Special Relativistic effects (like time dilation) in 4D. Other discussed applications included intelligent process control and multi-objective optimization.

Introduction to Formal Concept Analysis and Its Applications in Information Retrieval and Related Fields – Dmitry I. Ignatov, National Research University Higher School of Economics, Moscow, Russia

Introduced in the early 1980s by Rudolf Wille as a mathematical theory, Formal Concept Analysis (FCA) became a popular technique within the IR field. FCA is concerned with the

formalization of concepts and conceptual thinking and has been applied in many disciplines such as software engineering, machine learning, knowledge discovery, and data mining as well as ontology construction during the last 15-20 years. The underlying concept hierarchies extracted from objects-attributes data (like texts-keywords) and their line diagrams are suitable tools for visualization and exploration in IR; for instance, it was successfully applied to query refinement, document classification, ranking and browsing. This course offered a balanced combination of theoretical foundations, practice with main tools for FCA like Concept Explorer[6] and survey of relevant applications. Many of the used examples were real-life studies conducted by Dr. Ignatov.

Document Analysis and Retrieval in Scientific Digital Libraries: Case studies in applying Machine Learning for Information Retrieval – Sujatha Das G., Institute for Infocomm Research, Agency for Science and Technology Research, Singapore

The course presented by Dr. Das G. discussed the application of machine learning techniques in large-scale IR systems, namely digital libraries. This tutorial exploited document processing, retrieval and analysis tasks in CiteSeer, a digital library portal for scientific documents in Computer Science and related areas. The lecturer talked over various topics including: web crawling, document classification, content analysis with topic modeling tools, metadata extraction using sequential labeling, and ranking algorithms such as PageRank and HITS used in social and information network analysis.

Author Profiling and Plagiarism Detection – Paolo Rosso, Technical University of Valencia, Spain

The first part of the course demonstrated how author profiling helps in identifying personal characteristics including gender, age, native language, and even personality of the authors of a text. The second covered topic was plagiarism, i.e. re-usage of someone elses text, results, processes or prior ideas without explicitly acknowledging the original author. In the absence of the possibility to identify the sources of plagiarism, the detection of plagiarized fragments has to rely on unexpected irregularities through a document such as changes of style, vocabulary, or complexity. Paolo Rosso presented recent techniques for author profiling and plagiarism detection along with practical part, which featured experimentation with relevant text collections.

Sponsoring organizations made two scientific presentations additionally to the school program. Ludmila Ostroumova (Yandex) presented an overview of efficient approaches for crawling and indexing newly created web pages such as news, blog and forum posts. Dmitry Solovyov (Mail.Ru) gave a talk on application of Self-Organizing Maps to search engines analytics.

3 Young Scientist Conference

For the 8th time RuSSIR Young Scientist Conference was organized within the school program. The conference allowed to create a dialog between young researchers from different areas such as mathematics, computer science and linguistics as well as social and media sciences. The conference ran over two consecutive evenings and consisted of two parts, oral presentations and poster sessions.

[6] `http://conexp.sourceforge.net`

There were two types of submissions: full papers that underwent a thorough reviewing process and short poster notes. Out of 22 submitted full papers 8 were accepted for oral presentation at the conference and will be published in the school proceedings:

- Alexander Fonarev "Transformation of Categorical Features into Real Using Low-Rank Approximations"
- Rinat Gareev and Vladimir Ivanov "A Comparative Evaluation of Statistical Part-of-Speech Taggers for Russian"
- Alexandra Kaminskaya, Maria Mikhailova, Alexander Malioukov, and Dmitry Ignatov "Recommendation of Ideas and Antagonists for Crowdsourcing Platform Witology"
- Nikolay Karpov, Alexander Porshnev and Ilya Redkin "Modelling movement of stock market indexes with data from emoticons of Twitter users"
- Ksenia Konyushkova and Dorota Glowacka "ImSe: Exploratory Time-efficient Image Retrieval System"
- Andrey Kutuzov "Semantic Clustering of Russian Web Search Results: Possibilities and Problem"
- Galina Lezina and Pavel Braslavski "A Large-Scale Community Questions Classification Accounting for Category Similarity"
- Dmitry Ustalov "Towards Gamification and Cooperation in Linguistic Resources"

At the poster sessions all participants had an opportunity to discuss and exchange their research results and ideas. In total about 70 posters were displayed. As in the previous years the Young Scientist Conference was one of the main highlights of the school.

4 Social Programme

The RuSSIR Welcome Reception, served in cafe "Moloko" on the first day, August 18, provided participants with a perfect chance of getting to know each other. The participants were offered the possibility to play table-hockey, ping-pong and snooker as well to dance. On the second evening, the school participants were able to join bus and walking city tours in order to explore the magnificent Kremlin, to learn about city history, and to enjoy the beautiful views over the Volga river. Traditional sport event of open-air football and volleyball competition between students and teachers took place on Wednesday, August 20. The closing party on Thursday evening was held on the boat "Moskva 59" cruising on the Volga and Oka rivers.

5 School Proceedings

This year the first time RuSSIR proceedings are scheduled to be published in the Springer Communications in Computer and Information Science (CCIS) series[7]. The volume will feature two

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

sections: six tutorial notes ranging from 20 to 45 pages and eighth selected revised papers from the associated Young Scientist Conference up to 20 pages each.

6 Conclusions

The 8th Russian Summer School in Information Retrieval was a quite successful event: It brought together participants with diverse backgrounds from Russia and abroad and facilitated cross-disciplinary exchange of experience and ideas. Students had an unique opportunity to learn new material that is not usually presented in university curricula and got feedback from peers and teachers during the poster sessions and informal communications. The event contributed to supporting a lively IR community in Russia and establishing ties with international colleagues. We received very positive feedback from attendees on all the different aspects of the school.

The 9th RuSSIR will be held in Saint Petersburg in August 2015, and will be jointly organised by the Saint Petersburg chapter of the Higher School of Economics and ROMIP.

7 Acknowledgments

We thank all Local Organizing Committee members (namely, Alexey Malafeev, Dmitry Zelonkin, Cyril Sherstnyv, and Julia Baranova) for their commitment, which made the school possible, all Programme Committee members for their time and efforts ensuring a high level of quality for the RuSSIR 2014 programme and, in particular, all the lecturers and students who came to Nizhny Novgorod and made the school such a success. We also thank student volunteers who contributed to school organisation 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: National Research University Higher School of Economics[8] (main organizer), Yandex[9] and Mail.Ru[10] (golden level), Google[11] and ABBYY[12] (bronze level). We also thank the ELIAS network[13] of the European Science Foundation, and Springer representatives, namely Alfred Hofmann and Aliaksandr Birukou, for their support.

[8] http://www.hse.ru/
[9] http://yandex.com
[10] http://go.mail.ru/
[11] http://google.com/
[12] http://abbyy.com/
[13] http://www.elias-network.eu/