# 9th Russian Summer School in Information Retrieval (RuSSIR 2015)

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

Ilya Markov
University of Amsterdam,
The Netherlands
*i.markov@uva.nl*

Panos M. Pardalos
Universiry of Florida, USA &
National Research University
Higher School of Economics, Russia
*p.m.pardalos@gmail.com*

Yana Volkovich
Eurecat, Spain
*yana.volkovich@eurecat.edu*

Sergei Koltsov
National Research University
Higher School of Economics, Russia
*skoltsov@hse.ru*

Olessia Koltsova
National Research University
Higher School of Economics, Russia
*ekoltsova@hse.ru*

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

## 1 Introduction

The 9th Russian Summer School in Information Retrieval (RuSSIR 2015) was held on August 24-28, 2015 in St. Petersburg, 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 extensive international participation [1, 2]. Previously, RuSSIR took place in Yekaterinburg, Taganrog, Petrozavodsk, Voronezh, St. Petersburg, Yaroslavl, Kazan, and Nizhny Novgorod. RuS-

[1] http://romip.ru/russir2015/
[2] http://www.hse.ru/en/
[3] http://romip.ru/en/

SIR courses were taught by many prominent international researchers in Information Retrieval and related areas.

RuSSIR 2015 was held in St. Petersburg. St. Petersburg, the second largest city in Russia, was founded three centuries ago as the Russia's "window to Europe." Inspired by Venice and Amsterdam, tsar Peter the Great placed the new Russian capital at the beautiful delta of the Neva river on then the most western point of the country, and gave a start to erection of hundreds of its magnificent palaces that still remind of its imperial grandeur. After yielding the status of capital back to Moscow in 1918, St. Petersburg has become the Russia's major cultural center combining its fascinating Russian heritage with a distinctly European outlook.

The RuSSIR 2015 program featured courses focusing on social network analysis and graph mining along with traditional topics from Information Retrieval. The program consisted of two invited lectures, eight courses running in two parallel sessions, two sponsor talks, and the RuSSIR 2015 Young Scientist Conference.

The school welcomed 134 participants selected based on their applications. The majority of students came from Russia, but there were also 15 students from the European Union and seven from the rest of the world. The RuSSIR audience comprised of undergraduate, graduate, and doctoral students, as well as young academics and industrial developers. The total number of participants including students, sponsor representatives, lecturers and organizers was 169.

The participation was free of charge thanks to the sponsors. In addition, 20 accommodation grants were awarded to Russian participants by the Higher School of Economics and 14 European-based students received travel support from the European Science Foundation (ESF)[4] through the ELIAS network.[5]

## 2 Courses

The RuSSIR program was compiled based on submitted course proposals, reviewed by the RuSSIR Program Committee. Each course proposal was reviewed by at least four PC members. In total, 15 course proposals were submitted, five of which were selected for the school program. Additionally, there were five invited courses on the main topic of RuSSIR 2015, i.e., social network analysis and graph mining. Overall, the school program consisted of two plenary courses and eight regular courses run in two parallel tracks. Below, a brief overview of each course is given.

**Data Science for Massive (Dynamic) Networks**, Panos M. Pardalos (University of Florida, USA)

Data science tools, such as data mining and optimization heuristics, have been used to analyze many large (and massive) data-sets that can be represented as a network. In these networks, certain attributes are associated with vertices and edges. This analysis often provides useful information about the internal structure of the datasets they represent. The course presented the author's work on several networks from telecommunications (call graph), financial networks (market graph), social networks, and neuroscience.

**Community Detection in Networks**, Santo Fortunato (Aalto University, Finland)

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

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

The course was focused on one of the most popular topics in the network science: detection of communities in networks. Communities are usually conceived as subgraphs of a network, with a high density of links within the subgraphs and a comparatively lower density between them. The existence of community structure indicates that the nodes of the network are not homogeneous but divided into classes, with a higher probability of connections between nodes of the same class than between nodes of different classes. This can be due to various reasons. In a social network, for instance, communities could be groups of people with common interests, or acquaintanceships; in protein interaction networks they might indicate functional modules where proteins with the same function frequently interact in the cell, hence they share more links; in the web graph, they might be web pages dealing with similar topics, which therefore referring to each other.

**Text Quantification**, Fabrizio Sebastiani (Qatar Computing Research Institute, Qatar)

In a number of applications involving text classification in recent years it has been pointed out that the final goal is not determining which class (or classes) individual unlabeled documents belong to, but determining the prevalence (or "relative frequency") of each class in the unlabeled data. The latter task is known as text quantification (or prevalence estimation, or class prior estimation). The goal of this course was to introduce the audience to the problem of quantification, techniques that have been proposed for solving it, metrics used to evaluate them, applications in fields such as information retrieval, machine learning, and data mining, and to the open problems in the area.

**Leveraging Knowledge Graphs for Web Search**, Gianluca Demartini (University of Sheffield, UK)

Knowledge Graphs (KGs) contain structured information about entities such as persons, locations, and organizations. Modern Web Search engines leverage such KGs to empower entity-oriented search by displaying in search engine result pages so called entity cards that summarize the main facts about the queried entity. In this course, the author introduced the main concepts around KGs and the 'Web of data' and discussed techniques for mining the Web for entities and using KGs to create an entity-centric user experience on the Web. The author covered the Linked Open Data initiative including popular KGs such as Freebase, DBPedia, and Wikidata, introduced the Named Entity Recognition and Linking techniques, and discussed their usage for identifying entity mentions in textual content, disambiguating these mentions, and connecting them to entities in a background KG. Furthermore, the author presented techniques for entity search and micro-task crowdsourcing.

**Online/Offline Evaluation of Search Engines**, Evangelos Kanoulas (University of Amsterdam, The Netherlands)

Evaluation has played a critical role in the success of IR. There is an arsenal of methods in hand that researcher and practitioners use to evaluate an experimental search system and compare it to the production system. This course focused on the two predominant paradigms: collection-based evaluation and in-situ evaluation. Collection-based evaluation is performed offline, in a laboratory setting, while in-situ evaluation is run online, by deploying an experimental system and running user queries both against the experimental and the production system. The course covered the latest advances in both paradigms.

The course topics included click-models and model-based measures, measures for complex retrieval scenarios, statistical inference frameworks that allow hypothesis testing in complex ex-

perimental designs, and the state-of-the-art A/B testing and interleaving methods. Special focus was given to recent work that attempts to bridge the gap between these two evaluation paradigms, e.g., methods to predict the results of an A/B and interleaving test from offline historical data, collection-based evaluation frameworks with a human in the loop, etc.

**Models of Random Graphs and their Applications to the Web-graph Analysis**, Andrey Raigorodsky (Moscow Institute of Physics and Technology, Moscow State University, and Yandex, Russia)

This course provided an overview of various models for random graphs and their applications to the Web graph. The author started with the classical Erdös-Rényi model and its application to network reliability, then proceeded with the most recent models describing the topology and growth of the Internet, social networks, economic network, and biological networks, and finally presented several applications of these models to the problems of search and crawling.

**Visual object recognition and localization**, Ivan Laptev (INRIA Paris-Rocquencourt, France)

The goal of this course was to introduce state-of-the-art methods for large scale image recognition and retrieval. The course contained lectures and practical sessions. The lectures covered recent image representations for object recognition (HOG, SIFT, DPM, BOF, CNN) as well as modern machine learning techniques (SVM, CNN/Deep Learning). Besides lectures, the course included guided practical sessions where students were able to implement basic techniques for object recognition. As a result of the course, the participants learned about techniques enabling efficient search of particular object instances among billions of images. The participants also learned about most recent advances in Deep Learning enabling close-to-human performance for such tasks as face recognition and object category recognition.

**Contextual Search and Exploration**, Charles L. A. Clarke (University of Waterloo, Canada), Jaap Kamps (University of Amsterdam, The Netherlands), Julia Kiseleva (Eindhoven University of Technology, The Netherlands)

The ubiquitous availability of information on the web and personalized (mobile) devices has a revolutionary impact on modern information access, challenging both research and industrial practice. Searchers with a complex information need typically slice-and-dice their problem into several queries and subqueries, and laboriously combine the answers post hoc to solve their tasks. Rich context allows far more powerful, personalized search, without the need for users to write long complex queries.

This course discussed the challenges of contextual search and recommendation, with concrete focus on the venue recommendation task as run at TREC 2012-2015. It featured both lectures and hands-on sessions with data derived from the TREC task. The course enabled students to understand the challenges and opportunities of contextualized search over entities, to learn effective approaches to venue recommendation, and to get the hands-on experience with developing and evaluating personalized search and recommendation approaches.

**Big Data driven Logistics**, Athanasios Migdalas (Luleå University of Technology, Sweden)

The purpose of this course was to give an overview of a recent and very active field of Big Data Analytics (also referred to as Data Mining) with the focus on its application to Logistics and Supply Chain Management. The course covered basic and advanced Data Analytics methods, techniques related to Supply Chains, corresponding metrics, technologies and tools, and a number of research examples.

**Big Data Analytics with R**, Athanasia Karakitsiou (Luleå University of Technology, Sweden)

This course was a follow up of and a complement to the course "Big Data Driven Logistics." It offered an introductory guide to algorithms for Big Data Analytics using the R language. The course discussed time series decomposition and correlation, forecasting strategies, linear and non-linear regression and classification, clustering, etc. In particular, a number of methods related to Data Mining were considered, e.g., ARMA, ARIMA, Support Vector Machines (SVM), Naïve Bayes, Neural Networks, Discriminant Analysis, k-Center, etc.

Sponsoring organizations made three scientific presentations in addition to the main school program. Yandex presented a novel methodology that allows less exhaustive online experimentation for search engines; Mail.ru focused on user behavior analysis; Ok.ru discussed the connection between the size of data and its usefulness.

**Efficient Online Experiments**, Eugene Kharitonov (Yandex)

Online evaluation methods, such as A/B and interleaving experiments, are widely used for search engine evaluation in the industrial setting. Since they rely on noisy implicit user feedback, running each experiment takes a considerable time and a share of the query traffic. How to overcome this limitation is an important scientific and industrial problem.

This presentation discussed two recently proposed approaches to reducing the duration of the online evaluation experiments. First, the speaker explained how the sequential statistical testing methods can be applied in the online evaluation scenario. Such sequential testing procedures allow an experiment to stop early, once the data collected is sufficient to make a conclusion. Second, the lecturer presented a new modification of the widely used Team Draft interleaving algorithm, Generalized Team Draft. Generalized Team Draft achieves a faster convergence rate by performing a joint data-driven optimization of interleaving parameters, as well as by using a stratified estimate of the experiment outcome.

**Does the size matter? Smart Data at OK.ru**, Dmitry Bugaychenko (Odnoklassniki – OK.ru)

"Big data" is one of the top buzzwords in the IT world, but is it really about the size? The speaker discussed a set of cases from OK.ru where not so big data and a set of well tuned algorithms were used to significantly improve their services. He first considered several behavioral analyses and collaborative filtering algorithms and showed how domain knowledge might help to improve a recommender system. Then the speaker moved to the latent semantic mining algorithms and demonstrated how to generalize them not only to text mining, but also to mining semantics from user behavior. The speaker also described the infrastructure and approach used to turn well-known "slow" algorithms into real-time ones.

**Learning to Rank Using Clickthrough Data**, Vladimir Gulin (Search@Mail.Ru)

Which web page does a user want to retrieve when he/she types a query into a search engine? There are millions of pages that contain words from a user's query, but only a small subset of these pages are interesting to the user. If one knew the set of pages actually relevant to the user's query, one could use this as training data for optimizing the ranking formula. In this lecture a speaker proposed to collect such relevant pages based on implicit user feedback and discussed how to apply this approach in a commercial search engine.

# 3 Young Scientist Conference

For the 9th time the 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.

There were two types of submissions: full papers that underwent a thorough reviewing process and short poster notes. Out of 17 submitted full papers, 6 were accepted for oral presentation at the conference and will be published in the school proceedings:

- Ekaterina Pronoza, Elena Yagunova and Anton Pronoza. Construction of a Russian Paraphrase Corpus: Unsupervised Paraphrase Extraction;
- Julia Efremova, Alejandro Montes Garcia, Alfredo Bolt Iriondo and Toon Calders. Who are My Ancestors? Retrieving Family Relationships from Historical Texts;
- Alexey Raskin. Using Levenshtein Distance for Typical Users Actions Detection and Search Engine Switching Detection;
- Arshad Khan, Thanassis Tiropanis and David Martin. Exploiting Semantic Annotation of Content with Linked Open Data (LoD) to Improve Searching Performance in Web Repositories of Multi-disciplinary Research Data;
- Lyudmila Zaydelman, Irina Krylova, Boris Orekhov and Ekaterina Stepanova. Languages of Russia: Using Social Networks to Collect Texts;
- Igor Zakhlebin, Aleksandr Semenov, Alexander Tolmach and Sergey Nikolenko. Detecting Opinion Polarisation on Twitter by Constructing Pseudo-bimodal Networks of Mentions and Retweets.

During the poster sessions all participants had an opportunity to discuss and exchange their research results and ideas. As in the previous years, the Young Scientist Conference was one of the main highlights of the school.

# 4 Hackaton

Charles Clarke, Jaap Kamps and Julia Kiseleva organized a hackathon as an additional part of their tutorial. The hackathon was designed in a way similar to the TREC 2015 Contextual Suggestion Track.[6] The participants were asked to recommend to the lecturers a number of places to visit in St. Petersburg, based on the lecturers' profiles and external sources about the city.

The hackathon attracted 30 students who formed 10 teams. The winning teams were selected based on the originality, relevance, and efficiency of the proposed solutions:

[6] https://sites.google.com/site/treccontext/

- MAD IT (Best System Award) – Maria Zagulova, Andrey Poletaev, Dmitry Zhelonkin, Ivan Grechikhin, Tatiana Nikulina;
- SalsaRoulette (Best Presentation Award) – Navid Rekabsaz, Larisa Adamyan, Ioanna Miliou, Aldo Lipani;
- sleep_deprived (Most Original Approach Award) – Sagun Pai, Sheikh Muhammad Sarwar.

The lecturers visited several of the recommended venues over the weekend and confirmed that the suggestions were indeed highly relevant.

# 5 Social Program

RuSSIR 2015 was accompanied by two social events. The welcome party was held at the HSE premises on the first day of the school. It was aimed to give participants and lecturers an opportunity to meet each other in an informal environment. Also, during the welcome party the RuSSIR sponsors had a chance to present their companies. The RuSSIR party was held in the form of a boat trip on the fourth day of the school. The boat took the participants along the beautiful Neva river and its numerous branches to show the magnificent view of the old imperial St. Petersburg by night.

# 6 School Proceedings

This year it is the second time that 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 ranging from 20 to 45 pages and six selected revised papers from the associated Young Scientist Conference (up to 20 pages each). The previous proceedings are published in the CCIS vol. 505 [3].

# 7 Conclusions

The 9$^{th}$ Russian Summer School in Information Retrieval was a successful event: It brought together participants with diverse backgrounds from Russia and abroad and facilitated cross-disciplinary exchange of experience and ideas. Students had a unique opportunity to learn new material that is not usually present in university curricula and got feedback from 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. We received very positive feedback from attendees on all aspects of the school.

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

# 8 Acknowledgments

We thank all the local Organizing Committee members (especially, Daria Yudenkova) for their commitment, which made the school possible, all the Program Committee members for their time and efforts ensuring a high level of quality for the RuSSIR 2015 program and, in particular, all the lecturers and students who came to St. Petersburg and made the school such a success. We also thank student volunteers who contributed to the 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: National Research University Higher School of Economics[8] (main organizer), golden sponsors: Yandex,[9] Mail.Ru,[10] and the ELIAS network[11] of the European Science Foundation, bronze sponsors: Google,[12] Rambler,[13] JetBRAINS,[14] and the Russian Foundation for Basic Research.[15] We also grateful to Springer representatives, namely Alfred Hofmann and Aliaksandr Birukou, for their support.

# 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)

[3] 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

[8] http://www.hse.ru/
[9] http://yandex.com
[10] http://go.mail.ru/
[11] http://www.elias-network.eu/
[12] http://google.com/
[13] http://rambler-co.ru/en/
[14] http://www.jetbrains.com/
[15] http://www.rfbr.ru/rffi/eng