Research Networking Programmes

Science Meeting – Scientific Report

Russian School on Information Retrieval 2014

Application Reference N°: 5575

1) Summary

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

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.

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 the 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 organizers was 119.

School participation was free of charge thanks to the sponsorship support. In addition, 20 accommodation grants were awarded to Russian participants. Nine European-based students received travel support from the European Science

Foundation (ESF through the ELIAS network Travel expenses of three school teachers from Europe were also funded through the ELIAS/ESF grant.

2) Description of the scientific content of and discussions at the event

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
- Multimedia Analysis and Multimedia Visualization -- Marcel Worring, University of Amsterdam, The Netherlands
- Large Scale Information Retrieval -- Katja Hofmann, Microsoft Research, Cambridge, UK
- Web as a Corpus: Going Beyond the n-gram -- Preslav Nakov, Qatar Computing Research Institute, Doha, Qatar
- Visualization and Data Mining for High Dimensional Data -- Alfred Inselberg, Tel Aviv University, Israel and Pei Ling Lai, University of Science and Technology, Tainan, Taiwan
- 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
- 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
- Author Profiling and Plagiarism Detection -- Paolo Rosso, Technical University of Valencia, Spain
- Crawling of New Web Pages. Ludmila Ostroumova, Yandex
- Application of Self-Organizing Maps for Search Engine, Dimitry Solovyov, Mail.ru

3) Assessment of the results and impact of the event on the future directions of the field

The RUSSIR2014 has brought together a number of prominent lecturers covering a wide array of topics within the field of information retrieval where this year there was a particular emphasis on the role of visualization in Information Retrieval. This has been identified as an important element for the future of Information Retrieval. In all the courses the school made an important contribution to bringing the field forward especially by connecting the research in the EU with the highly skilled Russian students.

Annex 4a: Programme of the meeting

Seeking Simplicity in Search User Interfaces (SUI)

Marti Hearst

It is rare for a new user interface to break through and become successful, especially in information-intensive tasks like search, coming to consensus or building up knowledge. Most complex interfaces end up going unused. Often the successful solution lies in a previously unexplored part of the interface design space that is simple in a new way that works just right.

In this talk I will give examples of such successes in the information-intensive interface design space, and attempt to provide stimulating ideas for future research directions.

Bio: Marti Hearst is a Professor in the School of Information at UC Berkeley with an affiliate appointment in Computer Science. A primary focus of Dr. Hearst's research is user interfaces for search, and she is the author of the 2009 book Search User Interfaces. She has also researched extensively in computational linguistics and text mining with a focus on detecting semantic relations, and text segmentation including discourse boundaries and abbreviation recognition. Her more recent research interests include user interfaces for the exploratory text analysis in the digital humanities and peer learning in MOOCS. She was recently elected a Fellow of the ACM.

Multimedia Analysis and Multimedia Visualization (MAMV)

Marcel Worring

Tradionally Information Retrieval has an exclusive focus on textual information. Yet more and more information sources have images or videos as a prime information carrier. Such image collections are a tremendous source of information and knowledge. Yet due to the semantic gap it is difficult to get access to their content and properly employ their context such as tags, geolocation, and other metadata. To move forward, we propose multimedia analytics solutions. These methods bring human experts and machine algorithms together in a synergetic manner where the machine does the bulk processing and the expert deals with the difficult decisions. Advanced interactive visualizations provide the means to let the expert understand the results of automatic processing and interactively provide the system with further guidance. We present state-of-the-art semantic visual concept detectors which are able to see what is in the picture. From there we consider methods which learn the semantics of images from the interaction with the user in combination with advanced visualizations. Finally we consider methods for interactively summarizing and exploring the collection. In particular we propose multimedia pivot tables, an extension of the pivot table reports found in spreadsheets and show how they provide highly flexible and versatile explorative capabilities. They provide a step towards methods to get valuable insight in large image collections.

Bio: Marcel Worring is associate professor at the Intelligent Systems Lab Amsterdam.

He is co-initiator and associate director of the recently established Data Science Research Center Amsterdam bringing together the leading research institutes in Amsterdam (CWI, VU, UvA, HvA) on the whole data science chain from data acquisition, storage and distributed processing, to analysis, retrieval, and visualization. His research interests are in multimedia analytics, leveraging synergy in human-computer processes. He has published over 170 papers in refereed journals and conferences. He is general co-chair of ACM Multimedia 2016 in Amsterdam, was program co-chair for ACM Multimedia 2013 and ICMR 2013, and co-initiator and co-organizer of the VideOlympics 2007-2009. He was associate editor of Pattern Analysis and Applications and IEEE Transactions on Multimedia and currently is associate editor of ACM Transactions on Multimedia. He was guest editor for various special issues. He is senior member of both ACM and IEEE. He was a visiting researcher at Yale University (1992) and University of California San Diego (1998).

Online Experimentation for Information Retrieval (OEIR)

Katja Hofmann (katja.hofmann[a]microsoft.com)

Online experimentation focuses on insights that can be gained from user interactions with information retrieval (IR) systems. The most common form of online experimentation, A/B testing, is widely used, and has helped sustain continuous improvement of these systems. In research, online experimentation techniques can support naturalistic studies that draw conclusions based on users' natural interactions with experimental systems. As online experimentation has taken a more and more central role in a wide range of IR systems, most notably web search engines, a wide variety of techniques have been developed to improve the delity and sensitivity of results. For example, news recommendation and ad placement can now successfully adapt to users' preferences using so-called contextual bandit techniques. Preferences for search result rankings can be inferred using interleaved comparison techniques. Finally, online learning approaches are developed to automate part of the online experimentation, so that systems can continuously learn and adapt to their users.

This course discusses established and recently developed online experimentation techniques, including A/B testing, bandit approaches, interleaving, counterfactual analysis, and online learning to rank. Students learn to reason about the advantages and limitations of these methods, and learn to design online experiments. Finally, a practical component walks students through the process of implementing their own experiments and analyzing obtained results.

Web as a Corpus: Going Beyond the n-gram (WCGBn)

Preslav Nakov

The tutorial will offer 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 will be put on Web-based approaches that go beyond the n-gram. We will also discuss the reliability and stability of page hits when used as a proxy for n-gram frequencies. Several applications of these ideas will be 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 (VDMHDD)

Alfred Inselberg

Pei Ling Lai

The goal of visualization is to systematically incorporate our fantastic pattern-recognition into the problem-solving process. With parallel coordinates the perceptual barrier imposed by our 3-dimensional habitation is breached enabling the visualization of multidimensional problems. The highlights, from the foundations to the most recent results leading to the visualization of Big Data, are intuitively developed. By learning to discover patterns from the displays a powerful knowledge discovery process (USA patent) has evolved. It is illustrated on real multivariate datasets together with guidelines for exploration and good query design; relational patterns are more compact and informative than the data. Realizing that this approach is intrinsically limited leads to a deeper geometrical insight, the recognition of M-dimensional objects recursively from their (M − 1)-dimensional subsets. A linear N-dimensional relation is represented by (N − 1) indexed points. For example in 3-D, two points with two indices represent a line while two points with three indices represent a plane. Powerful geometrical algorithms (for intersections, con tainment, proximities) as well as applications including classification and collision avoidance emerge. A smooth surface in 3-D is the envelope of its tangent planes each of which is represented by 2 points with 3 indices. Hence, a 3-D surface can be represented by two planar regions (each consisting of the points with same indices) and in N-dimensions by (N − 1) planar regions. This is equivalent to representing a surface by its normal vectors (i.e. tangent planes) at each point. From these regions the surface properties in any dimension are revealed like, developable, convex, non-orientable (as the Mobius strip) yielding stunning patterns unlocking new geometrical insights. Non-convexities like folds, bumps, coiling, dimples and more are no longer hidden and are detected from just one or ientation. Evidently this representation is preferable for some applications like ray-tracing even in 3 -D. Such results were first discovered visually in 3-D and then proved mathematically for the higher dimensi ons; in the true spirit of Geometry! The patterns persist in the presence of errors (USA patent) and t hat is good news for the applications. Applications to collision avoidance for air traffic control (3 USA patents), non-linear interactive models of complex systems like a country’s economy and decision support for intensive care units, as well as Special Relativistic effects (like time dilation) in 4-D will be presented. Other applications include intelligent process control and multi-objective optimization.

Introduction to Formal Concept Analysis and Its Applications in Information Retrieval and Related Fields (FCAIR)

Dmitry I. Ignatov

Formal Concept Analysis (FCA) was introduced in the early 1980s by Rudolf Wille as a mathematical theory and is 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 ontology construction during the last 15-20 years. The core contributions of this course from IR perspective are based on our paper (Poelmans et al., 2012). This course is a balanced combination of theoretical foundations, practice and relevant applications: we provide intro to FCA, practice with main tools for FCA, FCA in Machine Learning and Data Mining, FCA in Information Retrieval and Text Mining, FCA in Ontology Modeling and other selected applications.

The FCA intro will be given at the solid and comprehensible level accessible even for newcomers. Many of the used examples are real-life studies conducted by the course author.

Document Analysis and Retrieval in Scientific Digital Libraries: Case studies in applying Machine Learning for Information Retrieval (DARSDL)

Sujatha Das G. (gsdas[a]cse.psu.edu)
Cornelia Caragea (ccaragea[a]unt.edu)
Xiaoli Li (xlli[a]i2r.a-star.edu.sg)
C. Lee Giles (giles[a]ist.psu.edu)

We will discuss the application of machine learning techniques in large-scale IR systems using digital libraries as representative systems. Digital library portals provide various IR applications for focused repositories of digital objects such as documents, video, and images. This course will be based on our experience with document processing, retrieval and analysis tasks in CiteSeer, a digital library portal for scientific documents in Computer Science and related areas [1]. We will discuss 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. Each self-contained course lecture has three parts: (1) A review of related IR concepts and models, (2) A presentation on sample state-of-the-art applications illustrating the discussed theory, and (3) A guided, hands-on exercise for attendees using publicly-available IR and data mining tools on large document collections obtained from well-known digital library portals.

Author Profiling and Plagiarism Detection (APPD)

Paolo Rosso; Natural Language Engineering Lab

Author profiling distinguishes between classes of authors studying how language is shared by classes of people. This task helps in identifying profiling aspects such as gender, age, native language, or even personality type. Author profiling is a problem of growing importance because allows, for instance, to spot potential online pedophiles analyzing their writing style and unveiling their age. From a forensic perspective it could be important, for instance, to know the linguistic profile of the author of a harassing text message (language used by a certain type of people) and identify certain characteristics (language as evidence).
Text re-use is the activity whereby pre-existing written texts are used again to create a new text or version of it. In case the re-use of someone else's text, results, processes or prior ideas occurs - without explicitly acknowledging the original author and source - then we talk about plagiarism. Not always it is possible to retrieve the source document(s) where plagiarism has been committed from (evidence of plagiarism). When this is not possible, the detection of plagiarized fragments has to rely on changes in the writing style in the analyzed document. The aim is detecting text inconsistencies (language as evidence), i.e., unexpected irregularities through a document such as changes of style, vocabulary, or complexity (e.g. writing style and complexity in texts vary with the author's age).

Crawling of New Web Pages

Ludmila Ostroumova; Yandex

Nowadays, more and more people use the Web as their primary source of up-to-date information. In this context, fast crawling and indexing of newly created Web pages has become crucial for search engines, especially because user traffic to a significant fraction of these new pages (like news, blog and forum posts) grows really quickly right after they appear, but lasts only for several days. I will talk about the problem of timely finding and crawling of such ephemeral new pages (in terms of user interest). Traditional crawling policies do not give any particular priority to such pages and may thus crawl them not quickly enough, and even crawl already obsolete content. Links to new pages which should be crawled can be found on older pages. In order to find them we need to periodically recrawl some old pages. One of the main difficulties here is to divide resources between these two activities, crawling and recrawling, in an efficient way. I will also discuss a metric, well thought out for this task, which takes into account the decrease of user interest for ephemeral pages over time. Another important problem here is predicting the page's popularity using only the features of the URL available at the time of its discovery. In addition to the total popularity of web pages, we also predicted the rate of popularity decay.

Application of Self-Organizing Maps for Search Engine

Dmitry Solovyov; Mail.ru

What is a search engine? I think many students will answer that they understand what SE is. But in real life SE has a very large infrastructure with different aspects. I‘d like to tell you about one of these SE features, that is, a data analysis for search engine development with Self-Organizing Maps.

The self-organizing map (SOM) was introduced in 1984 by Teuvo Kohonen and is a special type of neural networks. SOM implements projection of high dimensional space to low dimensional space and is an efficient tool for visualization of multidimensional numerical data. I’ll present some methods for visualizing SOM and describe how these methods can be used to analyze search engine data.

Annex 4b: Full list of speakers and participants

List of ELIAS funded lecturers

Paolo Rosso (Spain)
Katja Hoffman (UK)
Marcel Worring (NL)

List of ELIAS funded students

Jain, Arpit (France/India)
Cosma, Cristian Alexandru (Romania)
Jimborean, Ioana (Romania)
Lipani, Aldo (Austria)
Mehrotra, Rishabh (United Kingdom)
Paramonov, Sergey (Belgium)
Suciu, Darius Andrei (Romania)
Tolkacheva, Anna (the Netherlands)
Verma, Manisha (United Kingdom)