

Communications in Computer and Information Science

505

Commenced Publication in 2007

Founding and Former Series Editors:

Alfredo Cuzzocrea, Dominik Ślęzak, and Xiaokang Yang

Editorial Board

Simone Diniz Junqueira Barbosa

*Pontifical Catholic University of Rio de Janeiro (PUC-Rio),
Rio de Janeiro, Brazil*

Phoebe Chen

La Trobe University, Melbourne, Australia

Xiaoyong Du

Renmin University of China, Beijing, China

Joaquim Filipe

Polytechnic Institute of Setúbal, Setúbal, Portugal

Orhun Kara

TÜBİTAK BİLGEM and Middle East Technical University, Ankara, Turkey

Igor Kotenko

*St. Petersburg Institute for Informatics and Automation of the Russian
Academy of Sciences, St. Petersburg, Russia*

Ting Liu

Harbin Institute of Technology (HIT), Harbin, China

Krishna M. Sivalingam

Indian Institute of Technology Madras, Chennai, India

Takashi Washio

Osaka University, Osaka, Japan

More information about this series at <http://www.springer.com/series/7899>

Pavel Braslavski · Nikolay Karpov
Marcel Worring · Yana Volkovich
Dmitry I. Ignatov (Eds.)

Information Retrieval

8th Russian Summer School, RuSSIR 2014
Nizhniy Novgorod, Russia, August 18–22, 2014
Revised Selected Papers

Editors

Pavel Braslavski
Ural Federal University
Yekaterinburg
Russia

Nikolay Karpov
National Research University Higher School
of Economics
Nizhniy Novgorod
Russia

Marcel Worryng
Intelligent Systems Laboratory
University of Amsterdam
Amsterdam
The Netherlands

Yana Volkovich
Barcelona Media Research Foundation
Barcelona
Spain

Dmitry I. Ignatov
National Research University Higher School
of Economics
Moscow
Russia

ISSN 1865-0929 ISSN 1865-0937 (electronic)
Communications in Computer and Information Science
ISBN 978-3-319-25484-5 ISBN 978-3-319-25485-2 (eBook)
DOI 10.1007/978-3-319-25485-2

Library of Congress Control Number: 2015951408

Springer Cham Heidelberg New York Dordrecht London
© Springer International Publishing Switzerland 2015

This work is subject to copyright. All rights are reserved by the Publisher, whether the whole or part of the material is concerned, specifically the rights of translation, reprinting, reuse of illustrations, recitation, broadcasting, reproduction on microfilms or in any other physical way, and transmission or information storage and retrieval, electronic adaptation, computer software, or by similar or dissimilar methodology now known or hereafter developed.

The use of general descriptive names, registered names, trademarks, service marks, etc. in this publication does not imply, even in the absence of a specific statement, that such names are exempt from the relevant protective laws and regulations and therefore free for general use.

The publisher, the authors and the editors are safe to assume that the advice and information in this book are believed to be true and accurate at the date of publication. Neither the publisher nor the authors or the editors give a warranty, express or implied, with respect to the material contained herein or for any errors or omissions that may have been made.

Printed on acid-free paper

Springer International Publishing AG Switzerland is part of Springer Science+Business Media
(www.springer.com)

Preface

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 Economics² 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, RuSSIRs took place in Yekaterinburg, Taganrog, Petrozavodsk, Voronezh, Saint Petersburg, Yaroslavl, and Kazan. RuSSIR courses were taught by many prominent international researchers in information retrieval (IR) and related areas.

The 2014 RuSSIR program featured a track of courses focusing on visualization for information retrieval along with other topics related to IR. The program led to fruitful discussions among participants coming from different domains and allowed students to learn cross-disciplinary competencies. The school program 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 students 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 attendees was 119 including students, sponsor representatives, lecturers, and organizers.

School participation was free of charge thanks to the sponsorship support. In addition, 20 accommodation grants were awarded to Russian participants and nine of the European-based students received travel support from the European Science Foundation (ESF)⁴ through the ELIAS network⁵. Travel expenses of three school teachers from Europe were also funded through the ELIAS/ESF grant.

The RuSSIR program was compiled based on reviewing of submitted course proposals by the Program Committee. Each course proposal was reviewed by at least six PC members. In total, 17 course proposals were submitted, six of which were selected for the school program. Additionally, there were two invited lectures. Each of the six courses consisted of five 90-minute lectures taught during five subsequent days. The invited lectures ran as plenary sessions, the other six courses ran in two parallel sessions.

¹ <http://romip.ru/russir2014/>

² <http://www.hse.ru/en/>

³ <http://romip.ru/en/>

⁴ <http://www.esf.org/>

⁵ <http://www.elias-network.eu/>

- 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, Southern Taiwan 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

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, and posts in blogs and forums. Dmitry Solovyov (Mail.Ru) gave a talk on application of self-organizing maps to search engines analytics.

For the eight time the RuSSIR Young Scientist Conference has been held within the school program. The conference provided a platform for 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 went the thorough reviewing process and short poster notes. Out of 22 submitted full papers, 8 papers were accepted for oral presentation at the conference and published in the school proceedings.

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 program.

The volume features two sections: six tutorial notes and eighth revised papers from the associated Young Scientist Conference.

The 8th Russian Summer School in Information Retrieval was a successful event: It brought together researchers and students with different background, and facilitated interdisciplinary ideas exchange. 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.

We thank all the local Organizing Committee members (namely, Alexey Malafeev, Dmitry Zelonkin, Cyril Sherstnev, and Julia Baranova) for their commitment, which made the school possible, all the Programme Committee members for their time and efforts ensuring a high level of quality for the RuSSIR 2014 program and, in particular, all the lecturers and students who came to Nizhny Novgorod and made the school such a success. We also thank the 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 the generous financial support from our sponsors: National Research University Higher School of Economics⁶ (main organizer), Yandex⁷ and Mail.Ru⁸ (gold level), Google⁹ and ABBYY¹⁰ (bronze level). We are also grateful to the ELIAS network¹¹ of the European Science Foundation, and Springer representatives, namely, Alfred Hofmann and Aliaksandr Birukou, for their support.

February 2015

Pavel Braslavski
Nikolay Karpov
Marcel Worring
Yana Volkovich
Dmitry I. Ignatov

⁶ <http://www.hse.ru/>

⁷ <http://yandex.com>

⁸ <http://go.mail.ru/>

⁹ <http://google.com/>

¹⁰ <http://abby.com/>

¹¹ <http://www.elias-network.eu/>

Organization

The conference was organized by the National Research University Higher School of Economics (Nizhniy Novgorod, Russia).

Program Committee Chairs

Pavel Braslavski	Ural Federal University/Kontur Labs, Russia
Marcel Worring	University of Amsterdam, The Netherlands

Organizing Chair

Nikolay Karpov	National Research University Higher School of Economics, Russia
----------------	---

Proceedings Chair

Dmitry I. Ignatov	National Research University Higher School of Economics, Russia
-------------------	---

Young Scientific Conference Program Committee Chairs

Yana Volkovich	Barcelona Media, Spain/Cornell Tech, USA
Svitlana Volkova	Johns Hopkins University, USA

Steering Committee

Pavel Braslavski	Ural Federal University/Kontur Labs, Russia
Nikolay Karpov	National Research University Higher School of Economics, Russia
Nikita Zhiltsov	Kazan Federal University, Russia
Dmitry Chalyy	Yaroslavl Demidov State University, Russia
Alexander Goncharov	CVisionLab, Russia
Maxim Gubin	Google, USA
Ksenia Rogova	Katholieke Universiteit Leuven, Belgium
Alexander Sychev	Voronezh State University, Russia
Natalia Vassilieva	HP Labs, Russia

Organizing Committee

Valery Kalygin	National Research University Higher School of Economics, Russia
Fedor Vitugin	National Research University Higher School of Economics, Russia

Julia Baranova	National Research University Higher School of Economics, Russia
Alexey Malafeev	National Research University Higher School of Economics, Russia
Dmitry Zelonkin	National Research University Higher School of Economics, Russia
Cyril Sherstnyv	National Research University Higher School of Economics, Russia

Program Committee

Paolo Boldi	Università degli Studi di Milano, Italy
Peter Brusilovsky	University of Pittsburgh, USA
Claudio Carpineto	Fondazione Ugo Bordoni, Italy
Sergey Chernov	Yandex, Russia
Nicola Ferro	University of Padua, Italy
Maxim Gubin	Google, USA
Matthias Hagen	Bauhaus University Weimar, Germany
Dmitry I. Ignatov	National Research University Higher School of Economics, Russia
Hideo Joho	University of Tsukuba, Japan
Robert Jäschke	Leibniz Universität Hannover, Germany
Jaap Kamps	University of Amsterdam, The Netherlands
Daan Odijk	University of Amsterdam, The Netherlands
Stefan Rürger	The Open University, UK
Konstantin Savenkov	Moscow State University, Russia
Tobias Schreck	University of Konstanz, Germany
Pavel Serdyukov	Yandex, Russia
Denis Turdakov	Institute for System Programming of RAS, Russia
Natalia Vassilieva	HP Labs, Russia
Alexey Voropaev	Mail.Ru, Russia
Ingmar Weber	Qatar Computing Research Institute, Qatar
Michael Yudelson	Carnegie Learning, Inc., USA

Young Scientist Conference Program Committee

Mikhail Ageev	Moscow State University/National Research University Higher School of Economics, Russia
Ismail Altıngövd	Middle East Technical University, Turkey
Paolo Boldi	Università degli Studi di Milano, Italy
Pavel Braslavski	Ural Federal University/Kontur Labs, Russia
Claudio Carpineto	Fondazione Ugo Bordoni, Italy
Sergey Chernov	Yandex, Russia
Nicola Ferro	University of Padua, Italy
David Gleih	Purdue University, USA
Maxim Gubin	Google, USA

Matthias Hagen	Bauhaus University Weimar, Germany
Dmitry I. Ignatov	National Research University Higher School of Economics, Russia
Robert Jäschke	Leibniz Universität, Germany
Jaap Kamps	University of Amsterdam, The Netherlands
Nattiya Kanhabua	Leibniz Universität, Germany
Nikolay Karpov	National Research University Higher School of Economics, Russia
Maxim Khalilov	bmmt GmbH, Germany
Natalia Konstantinova	University of Wolverhampton, UK
Daan Odijk	University of Amsterdam, The Netherlands
Alexander Panchenko	Université catholique de Louvain, Belgium
Stefan Rüger	The Open University, UK
Pavel Serdyukov	Yandex, Russia
Denis Turdakov	Institute for System Programming of RAS, Russia
Antti Ukkonen	Aalto University, Finland
Natalia Vassilieva	HP Labs, Russia
Ingmar Weber	Qatar Computing Research Institute, Qatar
Marcel Worring	University of Amsterdam, The Netherlands
Michael Yudelson	Carnegie Learning, Inc., USA

Partners and Sponsoring Institutions

Gold sponsors	ABBYY	Yandex
Bronze sponsors	Google	Mail.Ru
Partners	ELIAS Network	European Science Foundation

Contents

Tutorial Papers

Document Analysis and Retrieval Tasks in Scientific Digital Libraries.	3
<i>Sujatha Das Gollapalli, Cornelia Caragea, Xiaoli Li, and C. Lee Giles</i>	
Online Experimentation for Information Retrieval	21
<i>Katja Hofmann</i>	
Introduction to Formal Concept Analysis and Its Applications in Information Retrieval and Related Fields	42
<i>Dmitry I. Ignatov</i>	
Visualization and Data Mining for High Dimensional Data: –With Connections to Information Retrieval.	142
<i>Alfred Inselberg and Pei Ling Lai</i>	
Web as a Corpus: Going Beyond the n -gram	185
<i>Preslav Nakov</i>	
Author Profiling and Plagiarism Detection	229
<i>Paolo Rosso</i>	

Young Scientists Conference Papers

Transformation of Categorical Features into Real Using Low-Rank Approximations	253
<i>Alexander Fonarev</i>	
A Comparative Evaluation of Statistical Part-of-Speech Taggers for Russian . . .	263
<i>Rinat Gareev and Vladimir Ivanov</i>	
Recommendation of Ideas and Antagonists for Crowdsourcing Platform Witology.	276
<i>Dmitry I. Ignatov, Maria Mikhailova, Alexandra Yu. Zakirova, and Alexander Malioukov</i>	
Modelling Movement of Stock Market Indexes with Data from Emoticons of Twitter Users	297
<i>Alexander Porshnev, Ilya Redkin, and Nikolay Karpov</i>	
ImSe: Exploratory Time-Efficient Image Retrieval System	307
<i>Ksenia Konyushkova and Dorota Glowacka</i>	

Semantic Clustering of Russian Web Search Results: Possibilities
and Problems 320
Andrey Kutuzov

A Large-Scale Community Questions Classification Accounting
for Category Similarity: An Exploratory Study 332
Galina Lezina and Pavel Braslavski

Towards Crowdsourcing and Cooperation in Linguistic Resources. 348
Dmitry Ustalov

Author Index 359