Lecture Notes in Computer Science 5766

Commenced Publication in 1973
Founding and Former Series Editors:
Gerhard Goos, Juris Hartmanis, and Jan van Leeuwen

Editorial Board

David Hutchison
Lancaster University, UK

Takeo Kanade
Carnegie Mellon University, Pittsburgh, PA, USA

Josef Kittler
University of Surrey, Guildford, UK

Jon M. Kleinberg
Cornell University, Ithaca, NY, USA

Alfred Kobsa
University of California, Irvine, CA, USA

Friedemann Mattern
ETH Zurich, Switzerland

John C. Mitchell
Stanford University, CA, USA

Moni Naor
Weizmann Institute of Science, Rehovot, Israel

Oscar Nierstrasz
University of Bern, Switzerland

C. Pandu Rangan
Indian Institute of Technology, Madras, India

Bernhard Steffen
University of Dortmund, Germany

Madhu Sudan
Microsoft Research, Cambridge, MA, USA

Demetri Terzopoulos
University of California, Los Angeles, CA, USA

Doug Tygar
University of California, Berkeley, CA, USA

Gerhard Weikum
Max-Planck Institute of Computer Science, Saarbruecken, Germany

Leif Azzopardi Gabriella Kazai
Stephen Robertson Stefan Rüger
Milad Shokouhi Dawei Song
Emine Yilmaz (Eds.)

Advances in Information Retrieval Theory

Second International Conference
on the Theory of Information Retrieval, ICTIR 2009
Cambridge, UK, September 10-12, 2009
Proceedings

Springer

Volume Editors

Leif Azzopardi
University of Glasgow
Department of Computing Science
Sir Alwyn Williams Building, Lilybank Gardens
Glasgow, G12 8QQ, Scotland, UK
E-mail: leif@dcs.gla.ac.uk

Gabriella Kazai
Stephen Robertson
Milad Shokouhi
Emine Yilmaz
Microsoft Research Ltd
7 J.J. Thomson Avenue, Cambridge, CB3 0FB, UK
E-mail: {gabkaz, ser, milads, eminey}@microsoft.com

Stefan Rüger
The Open University
Knowledge Media Institute
Milton Keynes, MK7 6AA, UK
E-mail: s.rueger@open.ac.uk

Dawei Song
The Robert Gordon University
School of Computing
St Andrew Street, Aberdeen, AB25 1HG, UK
E-mail: d.song@rgu.ac.uk

Library of Congress Control Number: 2009934008

CR Subject Classification (1998): H.3, H.2, I.2.3, I.2.6, F.2.2, H.4, H.5.2-4, I.7

LNCS Sublibrary: SL 3 – Information Systems and Application, incl. Internet/Web and HCI

ISSN 0302-9743
ISBN-10 3-642-04416-6 Springer Berlin Heidelberg New York
ISBN-13 978-3-642-04416-8 Springer Berlin Heidelberg New York

This work is subject to copyright. All rights are reserved, whether the whole or part of the material is concerned, specifically the rights of translation, reprinting, re-use of illustrations, recitation, broadcasting, reproduction on microfilms or in any other way, and storage in data banks. Duplication of this publication or parts thereof is permitted only under the provisions of the German Copyright Law of September 9, 1965, in its current version, and permission for use must always be obtained from Springer. Violations are liable to prosecution under the German Copyright Law.

springer.com

© Springer-Verlag Berlin Heidelberg 2009
Printed in Germany

Typesetting: Camera-ready by author, data conversion by Scientific Publishing Services, Chennai, India
Printed on acid-free paper SPIN: 12754809 06/3180 5 4 3 2 1 0

In memory of Sándor Dominich

Preface

These proceedings contain the refereed papers and posters presented at the Second International Conference on the Theory of Information Retrieval (ICTIR 2009), held at Microsoft Research in Cambridge, UK, September 10–11, 2009.

This biennial international conference provides an opportunity for the presentation of the latest work describing theoretical advances in the field of information retrieval (IR). The first ICTIR was held in Budapest in October 2007, organized by Keith van Rijsbergen, Sándor Dominich, Sándor Darányi, and Ferenc Kiss. ICTIR was brought about by the growing interest in the consecutive workshops run at ACM SIGIR each year from 2000 until 2005 on Mathematical and Formal Methods in IR (Athens, Greece, 2000; New Orleans, USA, 2001; Tampere, Finland, 2002; Toronto, Canada, 2003; Sheffield, UK, 2004; Salvador, Brazil, 2005). This sustained initiative was in a large part down to the determination of Sándor Dominich and his passion for all things good, formal and mathematical. The foundation and the success of ICTIR is a direct result of his commitment and dedication to fostering research and development into the theoretical underpinnings of IR. His dedication is epitomized by his two books on the subject: *Mathematical Foundations in Information Retrieval* published in 2001, and *The Modern Algebra of Information Retrieval* published in 2008. While his efforts to promoting formal methods for IR have led to the foundation of ICTIR, sadly, his untimely passing in 2008 means that he is unable to witness how the theory of IR unfolds in the future. Nonetheless, his belief in the importance of theory and his spirit in advocating the development of formal methods in IR lives on through this conference series. We dedicate ICTIR 2009 to Sándor Dominich as a tribute to his contribution to the field.

ICTIR 2009 presented the latest developments in IR and boasted a high-quality program covering a diverse range of topics. The papers accepted for publication and presentation at ICTIR 2009 were selected from a total of 82 submissions, which were received from Continental Europe (39%), UK (21%), North America (18%), Asia and Australasia (10%), Middle East and Africa (12%). The submissions were assessed by at least three reviewers in a double-blind review process, and were ranked according to their scientific quality, originality, and contribution to the theory of IR. In total, 18 full papers (22%), 14 short papers (17%), and 11 posters (13%) were accepted. We categorized the accepted contributions into four main themes: novel IR models, evaluation, efficiency, and new perspectives in IR. Twenty-one papers fall into the general theme of novel IR models, ranging from various retrieval models (8), query and term selection models (4), Web IR models (3), developments in novelty and diversity (3), to the modeling of user aspects (3). There are four papers on new evaluation methodologies, e.g., modeling score distributions, evaluation over sessions, and an axiomatic framework for XML retrieval evaluation. Three papers focus on the

issue of efficiency and offer solutions to improve the tractability of PageRank, data-cleansing practices for training classifiers, and approximate search for distributed IR. Finally, four papers look into new perspectives of IR and shed light on some new emerging areas of interest, such as the application and adoption of quantum theory in IR.

We would like to thank the invited speaker, Peter Bruza, for his thought-provoking keynote speech on using quantum theory to develop a new suite of information-processing models that are motivated from a cognitive science perspective. We would also like to thank all the authors who submitted their work for consideration, and all the participants and student volunteers for their contributions and help. We are grateful to the members of the Program Committee for their time and effort in providing timely and high-quality feedback and reviews.

Finally, we would like to say special thanks to the following organizations and individuals who helped to make ICTIR 2009 a success:

- Microsoft Research for hosting the event and providing the excellent conference facilities, as well sponsoring the conference dinner. We especially thank Rachael Billing (overall organization, banquet, catering), Nick Duffield (graphics design, marketing materials), Sarah Head (marketing, conference bags), Sarah Nightingale (facilities), Fabien Petitcolas (sponsorship), Mari Ann Lindqvist (finance), Adrian Cooper (security), Ian Kelly (IT) and the entire IT support team.
- The Open University for providing conference website design, registration and financial management. Many thanks go to Damian Dadswell (Web), Harriett Cornish (initial graphical designs), and Jane Whild, Rachel Barnett, Aneta Tumilowicz and The Open University's Finance devision (budget and financial management).
- The British Computer Society - Information Retrieval Specialist Group (BCS-IRSG) for providing financial support for students and for sponsoring 40 copies of book *The Modern Algebra of Information Retrieval* as the tribute to Sándor Dominich.
- The editorial staff at Springer for their agreement and assistance in publishing the conference as part of the *Lecture Notes in Computer Science* (LNCS) series.
- Yahoo Research for sponsoring the Best Student Paper Award.
- True Knowledge for their kind sponsorship.

September 2009

Leif Azzopardi
Gabriella Kazai
Stephen Robertson
Stefan Rüger
Milad Shokouhi
Dawei Song
Emine Yilmaz

Organization

Organizing Institutions

ICTIR 2009 was organized by Microsoft Research Cambridge, the Knowledge Media Institute of the Open University, the Department of Computing Science of University of Glasgow, and the School of Computing of the Robert Gordon University.

Conference Chairs

Conference Chairs	Gabriella Kazai, Microsoft Research, UK
	Stefan Rüger, The Open University, UK
Program Chairs	Leif Azzopardi, University of Glasgow, UK
	Dawei Song, The Robert Gordon University, UK
Honorary Chair	Keith van Rijsbergen, University of Glasgow, UK
Local Chairs	Stephen Robertson, Microsoft Research, UK
	Milad Shokouhi, Microsoft Research, UK
	Emine Yilmaz, Microsoft Research, UK

Sponsors

Microsoft Research
The Open University
British Computer Society - Information Retrieval Specialist Group (BCS-IRSG)
True Knowledge
Yahoo Research

Program Committee

Gianni Amati	FUB, Italy
Hany Azzam	Queen Mary University of London, UK
Richard Bache	University of Glasgow, UK
Jing Bai	Yahoo! Inc., USA
Mark Baillie	University of Strathclyde, UK
Roberto Basili	University of Rome Tor Vergata, Italy
Nick Belkin	Rutgers University, USA
Bodo Billerbeck	Microsoft Research, Cambridge, UK
Giorgio Brajnik	University of Udine, Italy
Peter Bruza	Queensland University of Technology, Australia
Di Cai	University of Glasgow, UK
Steven Cater	Kettering University, USA
Fabio Crestani	University of Lugano, Switzerland
Tamas Doszkocs	National Library of Medicine, USA
Hui Fang	University of Delaware, USA
Jan Frederik Forst	Queen Mary University of London, UK
Norbert Fuhr	University of Duisburg-Essen, Germany
Susan Gauch	University of Arkansas, USA
Thore Graepel	Microsoft Research, Cambridge, UK
Martin Halvey	University of Glasgow, UK
Claudia Hauff	University of Twente, The Netherlands
Ben He	University of Glasgow, UK
Djoerd Hiemstra	University of Twente, The Netherlands
Eduard Hoenkamp	Maastricht University, The Netherlands
Qiang Huang	The Robert Gordon University, UK
Jimmy Huang	York University, Canada
Theo Huibers	University of Twente, The Netherlands
Peter Ingwersen	Royal School of Library and Information Science, Denmark
Kalervo Jarvelin	Tampere University, Finland
Gareth Jones	Dublin City University, Ireland
April Kontostathis	Ursinus College, USA
Udo Kruschwitz	University of Essex, UK
Mounia Lalmas	University of Glasgow, UK
Wai Lam	Chinese University of Hong Kong, Hong Kong
Birger Larsen	Royal School of Library and Information Science, Denmark
Raymond Lau	City University of Hong Kong, Hong Kong
Victor Lavrenko	University of Edinburgh, UK
Christina Lioma	K.U. Leuven, Belgium

David Losada	Universidad de Santiago de Compostela, Spain
Robert Luk	Hong Kong Polytechnic University, Hong Kong
Massimo Melucci	University of Padua, Italy
Stefano Mizzaro	University of Udine, Italy
Dunja Mladenic	J. Stefan Institute, Slovenia
Nikolaos Nanas	Centre for Research and Technology - Thessaly, Greece
Jian-Yun Nie	University of Montreal, Canada
Paul Ogilvie	mSpoke, USA
Iadh Ounis	University of Glasgow, UK
Benjamin Piwowarski	University of Glasgow, UK
Filip Radlinski	Microsoft Research, Cambridge, UK
Vijay Raghavan	University of Louisiana at Lafayette, USA
Maarten de Rijke	University of Amsterdam, The Netherlands
Stephen Robertson	Microsoft Research, Cambridge, UK
Ian Ruthven	University of Strathclyde, UK
Milad Shokouhi	Microsoft Research, Cambridge, UK
Laurianne Sitbon	National ICT Centre, Australia
Edward Snelson	Microsoft Research, Cambridge, UK
Amanda Spink	Queensland University of Technology, Australia
Paul Thomas	CSIRO ICT Centre, Australia
Olga Vechtomova	University of Waterloo, Canada
Vishwa Vinay	Microsoft Research, Cambridge, UK
Jun Wang	The Robert Gordon University, UK
Jun Wang	University College London, UK
Wensi Xi	Google, USA
Emine Yilmaz	Microsoft Research, Cambridge, UK
Dell Zhang	Birkbeck, University of London, UK
Peng Zhang	The Robert Gordon University, UK
Jianhan Zhu	University College London, UK
Guido Zuccon	University of Glasgow, UK

Table of Contents

Invited Talk

Is There Something Quantum-Like about the Human Mental Lexicon? 1
Peter Bruza

Regular Papers

Efficiency

Probably Approximately Correct Search 2
Ingemar J. Cox, Ruoxun Fu, and Lars Kai Hansen

PageRank: Splitting Homogeneous Singular Linear Systems of Index One 17
Douglas V. de Jager and Jeremy T. Bradley

Training Data Cleaning for Text Classification 29
Andrea Esuli and Fabrizio Sebastiani

Retrieval Models

Semi-parametric and Non-parametric Term Weighting for Information Retrieval 42
Donald Metzler and Hugo Zaragoza

Bridging Language Modeling and Divergence from Randomness Models: A Log-Logistic Model for IR 54
Stéphane Clinchant and Eric Gaussier

Ordinal Regression Based Model for Personalized Information Retrieval 66
Mohamed Farah

Navigating in the Dark: Modeling Uncertainty in Ad Hoc Retrieval Using Multiple Relevance Models 79
Natali Soskin, Oren Kurland, and Carmel Domshlak

Query and Term Models

A Belief Model of Query Difficulty That Uses Subjective Logic 92
Christina Lioma, Roi Blanco, Raquel Mochales Palau, and Marie-Francine Moens

"A Term is Known by the Company It Keeps": On Selecting a Good Expansion Set in Pseudo-Relevance Feedback 104
Raghavendra Udupa, Abhijit Bhole, and Pushpak Bhattacharyya

An Effective Approach to Verbose Queries Using a Limited Dependencies Language Model 116
Eduard Hoenkamp, Peter Bruza, Dawei Song, and Qiang Huang

Time-Sensitive Language Modelling for Online Term Recurrence Prediction 128
Dell Zhang, Jinsong Lu, Robert Mao, and Jian-Yun Nie

Evaluation

Score Distributions in Information Retrieval 139
Avi Arampatzis, Stephen Robertson, and Jaap Kamps

Modeling the Score Distributions of Relevant and Non-relevant Documents 152
Evangelos Kanoulas, Virgil Pavlu, Keshi Dai, and Javed A. Aslam

Modeling Expected Utility of Multi-session Information Distillation 164
Yiming Yang and Abhimanyu Lad

Specificity Aboutness in XML Retrieval 176
Tobias Blanke and Mounia Lalmas

Novelty and Diversity

An Effectiveness Measure for Ambiguous and Underspecified Queries ... 188
Charles L.A. Clarke, Maheedhar Kolla, and Olga Vechtomova

An Analysis of NP-Completeness in Novelty and Diversity Ranking 200
Ben Carterette

From "Identical" to "Similar": Fusing Retrieved Lists Based on Inter-document Similarities 212
Anna Khudyak Kozorovitzky and Oren Kurland

Short Papers

New Perspectives in IR

A Quantum-Based Model for Interactive Information Retrieval 224
Benjamin Piwowarski and Mounia Lalmas

The Quantum Probability Ranking Principle for Information Retrieval 232
Guido Zuccon, Leif A. Azzopardi, and Keith van Rijsbergen

Written Texts as Statistical Mechanical Problem 241
Kostadin Koroutchev, Elka Korutcheva, and Jian Shen

What Happened to Content-Based Information Filtering? 249
Nikolaos Nanas, Anne De Roeck, and Manolis Vavalis

Retrieval Models

Prior Information and the Determination of Event Spaces in Probabilistic Information Retrieval Models 257
Corrado Boscarino and Arjen P. de Vries

Robust Word Similarity Estimation Using Perturbation Kernels 265
Kevyn Collins-Thompson

Possibilistic Similarity Estimation and Visualization 273
Anas Dahabiah, John Puentes, and Basel Solaiman

A New Measure of the Cluster Hypothesis 281
Mark D. Smucker and James Allan

User Aspects

Explaining User Performance in Information Retrieval: Challenges to IR Evaluation 289
Kalervo Järvelin

A Four-Factor User Interaction Model for Content-Based Image Retrieval 297
Haiming Liu, Victoria Uren, Dawei Song, and Stefan Rüger

Predicting Query Performance by Query-Drift Estimation 305
Anna Shtok, Oren Kurland, and David Carmel

Web Models

What's in a Link? From Document Importance to Topical Relevance ... 313
Marijn Koolen and Jaap Kamps

Avoiding Bias in Text Clustering Using Constrained K-means and May-Not-Links . . . 322
M. Eduardo Ares, Javier Parapar, and Álvaro Barreiro

Optimizing WebPage Interest . . . 330
Willem Elbers and Theo van der Weide

Posters

The "Beautiful" in Information: Philosophy of Aesthetics and Information Visualization . . . 338
Gerald Benoit

IR Evaluation without a Common Set of Topics . . . 342
Matteo Cattelan and Stefano Mizzaro

An Ad Hoc Information Retrieval Perspective on PLSI through Language Model Identification . . . 346
Jean-Cédric Chappelier and Emmanuel Eckard

Less is More: Maximal Marginal Relevance as a Summarisation Feature . . . 350
Jan Frederik Forst, Anastasios Tombros, and Thomas Roelleke

On the Notion of "An Information Need" . . . 354
Eduard Hoenkamp

A Logical Inference Approach to Query Expansion with Social Tags . . . 358
Christina Lioma, Roi Blanco, and Marie-Francine Moens

Evaluating Mobile Proactive Context-Aware Retrieval: An Incremental Benchmark . . . 362
Davide Menegon, Stefano Mizzaro, Elena Nazzi, and Luca Vassena

Predicting the Usefulness of Collection Enrichment for Enterprise Search . . . 366
Jie Peng, Ben He, and Iadh Ounis

Ranking List Dispersion as a Query Performance Predictor . . . 371
Joaquín Pérez-Iglesias and Lourdes Araujo

Semi-subsumed Events: A Probabilistic Semantics of the BM25 Term Frequency Quantification . . . 375
Hengzhi Wu and Thomas Roelleke

Batch-Mode Computational Advertising Based on Modern Portfolio Theory . . . 380
Dell Zhang and Jinsong Lu

Author Index . . . 385