

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

TU Dortmund University, 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

Shlomo Geva Jaap Kamps
Andrew Trotman (Eds.)

Focused Retrieval and Evaluation

8th International Workshop of the Initiative
for the Evaluation of XML Retrieval, INEX 2009
Brisbane, Australia, December 7-9, 2009
Revised and Selected Papers

Volume Editors

Shlomo Geva

Queensland University of Technology, Faculty of Science and Technology
GPO Box 2434, Brisbane Qld 4001, Australia

E-mail: s.geva@qut.edu.au

Jaap Kamps

University of Amsterdam, Archives and Information Studies/Humanities
Turfdraagsterpad 9, 1012 XT Amsterdam, The Netherlands

E-mail: kamps@uva.nl

Andrew Trotman

University of Otago, Department of Computer Science
P.O. Box 56, Dunedin 9054, New Zealand

E-mail: andrew@cs.otago.ac.nz

Library of Congress Control Number: 2010930655

CR Subject Classification (1998): H.3, H.3.3, H.3.4, H.2.8, H.2.3, H.2.4, E.1

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

ISSN 0302-9743

ISBN-10 3-642-14555-8 Springer Berlin Heidelberg New York

ISBN-13 978-3-642-14555-1 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 2010
Printed in Germany

Typesetting: Camera-ready by author, data conversion by Scientific Publishing Services, Chennai, India
Printed on acid-free paper 06/3180

Preface

Welcome to the proceedings of the 8th Workshop of the Initiative for the Evaluation of XML Retrieval (INEX)! Now in its eighth year, INEX is an established evaluation forum for XML information retrieval (IR), with over 100 organizations worldwide registered and over 50 groups participating actively in at least one of the tracks. INEX aims to provide an infrastructure, in the form of a large structured test collection and appropriate scoring methods, for the evaluation of focused retrieval systems.

XML IR plays an increasingly important role in many information access systems (e.g., digital libraries, Web, intranet) where content is a mixture of text, multimedia, and metadata, formatted according to the adopted W3C standard for information repositories, the so-called eXtensible Markup Language (XML). The ultimate goal of such systems is to provide the right content to their end-users. However, while many of today's information access systems still treat documents as single large (text) blocks, XML offers the opportunity to exploit the internal structure of documents in order to allow for more precise access, thus providing more specific answers to user requests. Providing effective access to XML-based content is therefore a key issue for the success of these systems.

INEX 2009 was an exciting year for INEX in which a new collection was introduced that is again based on Wikipedia but is more than four times larger, with longer articles and additional semantic annotation. In total, eight research tracks were included, which studied different aspects of focused information access:

Ad Hoc Track investigated the effectiveness of XML-IR and Passage Retrieval for four ad hoc retrieval tasks: Thorough, Focused, Relevant in Context, and Best in Context.

Book Track investigated information access to, and IR techniques for, searching full texts of digitized books.

Efficiency Track investigated both the effectiveness and efficiency of XML ranked retrieval approaches on real data and real queries.

Entity Ranking Track investigated entity retrieval rather than text retrieval: 1) Entity Ranking, 2) Entity List Completion.

Interactive Track investigated the behavior of users when interacting with XML documents, as well as developed retrieval approaches which are effective in user-based environments.

Question Answering Track investigated technology for accessing structured documents that can be used to address real-world focused information needs formulated as natural language questions.

Link the Wiki Track investigated link discovery between Wikipedia documents, both at the file level and at the element level.

XML Mining Track investigated structured document mining, especially the classification and clustering of structured documents.

The aim of the INEX 2009 workshop was to bring together researchers in the field of XML IR who participated in the INEX 2009 campaign. During the past year, participating organizations contributed to the building of large-scale XML test collections by creating topics, performing retrieval runs, and providing relevance assessments. The workshop concluded the results of this effort, summarized and addressed issues encountered, and devised a work plan for the future evaluation of XML retrieval systems. These proceedings report the final results of INEX 2009. We received 47 submissions, already being a selection of work at INEX, and accepted a total of 42 papers based on peer-reviewing, yielding a 89% acceptance rate.

All INEX tracks start from having available suitable text collections. We gratefully acknowledge the data made available by: Amazon (Interactive Track), New Zealand Ministry for Culture and Heritage (*Te Ara*, Link-the-Wiki Track), Microsoft (Book Track), Wikipedia, and Ralf Schenkel of the Max-Planck Institute for the conversion of the Wikipedia.

INEX has outgrown its previous home at *Schloss Dagstuhl* and was held in Brisbane, Australia. Thanks to Richi Nayak and the QUT team for preserving the unique atmosphere of INEX—a setting where informal interaction and discussion occur naturally and frequently—in the unique location of the Woodlands of Marburg. Thanks to HCSNet, the Australian Research Council’s Research Network in Human Communication Science, for sponsoring the invited talks by Peter Bruza (QUT), Cécile Paris (CSIRO), and Ian Witten (Waikato). Finally, INEX is run for, but especially by, the participants. It was a result of tracks and tasks suggested by participants, topics created by participants, systems built by participants, and relevance judgments provided by participants. So the main thank you goes to each of these individuals!

April 2010

Shlomo Geva
Jaap Kamps
Andrew Trotman

Organization

Steering Committee

Charlie Clarke	University of Waterloo, Canada
Norbert Fuhr	University of Duisburg-Essen, Germany
Shlomo Geva	Queensland University of Technology, Australia
Jaap Kamps	University of Amsterdam, The Netherlands
Mounia Lalmas	Queen Mary, University of London, UK
Stephen Robertson	Microsoft Research Cambridge, UK
Andrew Trotman	University of Otago, New Zealand
Ellen Voorhees	NIST, USA

Chairs

Shlomo Geva	Queensland University of Technology, Australia
Jaap Kamps	University of Amsterdam, The Netherlands
Andrew Trotman	University of Otago, New Zealand

Track Organizers

Ad Hoc

Shlomo Geva	Queensland University of Technology, Australia
Jaap Kamps	University of Amsterdam, The Netherlands
Miro Lethonen	University of Helsinki, Finland
Ralf Schenkel	Max-Planck-Institut für Informatik, Germany
James A. Thom	RMIT, Australia
Andrew Trotman	University of Otago, New Zealand

Book

Antoine Doucet	University of Caen, France
Gabriella Kazai	Microsoft Research Cambridge, UK
Marijn Koolen	University of Amsterdam, The Netherlands
Monica Landoni	University of Lugano, Switzerland

Efficiency

Ralf Schenkel	Max-Planck-Institut für Informatik, Germany
Martin Theobald	Max-Planck-Institut für Informatik, Germany

Entity Ranking

Gianluca Demartini	L3S, Germany
Tereza Iofciu	L3S, Germany
Arjen de Vries	CWI, The Netherlands

Interactive

Thomas Beckers	University of Duisburg-Essen, Germany
Nisa Fachry	University of Amsterdam, The Netherlands
Norbert Fuhr	University of Duisburg-Essen, Germany
Ragnar Nordlie	Oslo University College, Norway
Nils Pharø	Oslo University College, Norway

Link the Wiki

Shlomo Geva	Queensland University of Technology, Australia
Darren Huang	Queensland University of Technology, Australia
Andrew Trotman	University of Otago, New Zealand

Question Answering

Patrice Bellot	University of Avignon, France
Veronique Moriceau	LIMSI-CNRS, France
Eric SanJuan	University of Avignon, France
Xavier Tannier	LIMSI-CNRS, France

XML Mining

Shlomo Geva	Queensland University of Technology, Australia
Ludovic Denoyer	University Paris 6, France
Chris De Vries	Queensland University of Technology, Australia
Patrick Gallinari	University Paris 6, France
Sangeetha Kutty	Queensland University of Technology, Australia
Richi Nayak	Queensland University of Technology, Australia

Table of Contents

Invited

Is There Something Quantum-Like about the Human Mental Lexicon?	1
<i>Peter Bruza</i>	
Supporting for Real-World Tasks: Producing Summaries of Scientific Articles Tailored to the Citation Context	2
<i>Cécile Paris</i>	
Semantic Document Processing Using Wikipedia as a Knowledge Base	3
<i>Ian H. Witten</i>	

Ad Hoc Track

Overview of the INEX 2009 Ad Hoc Track	4
<i>Shlomo Geva, Jaap Kamps, Miro Lethonen, Ralf Schenkel, James A. Thom, and Andrew Trotman</i>	
Analysis of the INEX 2009 Ad Hoc Track Results	26
<i>Jaap Kamps, Shlomo Geva, and Andrew Trotman</i>	
ENSM-SE at INEX 2009 : Scoring with Proximity and Semantic Tag Information	49
<i>Michel Beigbeder, Amélie Imafouo, and Annabelle Mercier</i>	
LIP6 at INEX'09 : OWPC for Ad Hoc Track	59
<i>David Buffoni, Nicolas Usunier, and Patrick Gallinari</i>	
A Methodology for Producing Improved Focused Elements	70
<i>Carolyn J. Crouch, Donald B. Crouch, Dinesh Bhirud, Pavan Poluri, Chaitanya Polumetla, and Varun Sudhakar</i>	
ListBM: A Learning-to-Rank Method for XML Keyword Search	81
<i>Ning Gao, Zhi-Hong Deng, Yong-Qing Xiang, and Yu Hang</i>	
UJM at INEX 2009 Ad Hoc Track	88
<i>Mathias Géry and Christine Largeron</i>	
Language Models for XML Element Retrieval.....	95
<i>Rongmei Li and Theo van der Weide</i>	

Use of Language Model, Phrases and Wikipedia Forward Links for INEX 2009	103
<i>Philippe Mulhem and Jean-Pierre Chevallet</i>	
Parameter Tuning in Pivoted Normalization for XML Retrieval: ISI@INEX09 Adhoc Focused Task.....	112
<i>Sukomal Pal, Mandar Mitra, and Debasis Ganguly</i>	
Combining Language Models with NLP and Interactive Query Expansion	122
<i>Eric SanJuan and Fidelia Ibekwe-SanJuan</i>	
Exploiting Semantic Tags in XML Retrieval	133
<i>Qiyue Wang, Qiushi Li, Shan Wang, and Xiaoyong Du</i>	

Book Track

Overview of the INEX 2009 Book Track	145
<i>Gabriella Kazai, Antoine Doucet, Marijn Koolen, and Monica Landoni</i>	
XRCE Participation to the 2009 Book Structure Task	160
<i>Hervé Déjean and Jean-Luc Meunier</i>	
The Book Structure Extraction Competition with the Resurgence Software at Caen University	170
<i>Emmanuel Giguet and Nadine Lucas</i>	
Ranking and Fusion Approaches for XML Book Retrieval	179
<i>Ray R. Larson</i>	
OUC's Participation in the 2009 INEX Book Track.....	190
<i>Michael Preminger, Ragnar Nordlie, and Nils Pharo</i>	

Efficiency Track

Overview of the INEX 2009 Efficiency Track	200
<i>Ralf Schenkel and Martin Theobald</i>	
Index Tuning for Efficient Proximity-Enhanced Query Processing	213
<i>Andreas Broschart and Ralf Schenkel</i>	
TopX 2.0 at the INEX 2009 Ad-Hoc and Efficiency Tracks: Distributed Indexing for Top-k-Style Content-And-Structure Retrieval.....	218
<i>Martin Theobald, Ablimit Aji, and Ralf Schenkel</i>	
Fast and Effective Focused Retrieval.....	229
<i>Andrew Trotman, Xiang-Fei Jia, and Shlomo Geva</i>	

Achieving High Precisions with Peer-to-Peer Is Possible!	242
<i>Judith Winter and Gerold Kühne</i>	

Entity Ranking Track

Overview of the INEX 2009 Entity Ranking Track	254
<i>Gianluca Demartini, Tereza Iofciu, and Arjen P. de Vries</i>	
Combining Term-Based and Category-Based Representations for Entity Search	265
<i>Krisztian Balog, Marc Bron, Maarten de Rijke, and Wouter Weerkamp</i>	
Focused Search in Books and Wikipedia: Categories, Links and Relevance Feedback	273
<i>Marijn Koolen, Rianne Kaptein, and Jaap Kamps</i>	
A Recursive Approach to Entity Ranking and List Completion Using Entity Determining Terms, Qualifiers and Prominent n-Grams	292
<i>Madhu Ramanathan, Srikanth Rajagopal, Venkatesh Karthik, Meenakshi Sundaram Murugesan, and Saswati Mukherjee</i>	

Interactive Track

Overview of the INEX 2009 Interactive Track	303
<i>Nils Pharo, Ragnar Nordlie, Norbert Fuhr, Thomas Beckers, and Khairun Nisa Fachry</i>	

Link the Wiki Track

Overview of the INEX 2009 Link the Wiki Track	312
<i>Wei Che (Darren) Huang, Shlomo Geva, and Andrew Trotman</i>	
An Exploration of Learning to Link with Wikipedia: Features, Methods and Training Collection	324
<i>Jiyin He and Maarten de Rijke</i>	
University of Waterloo at INEX 2009: Ad Hoc, Book, Entity Ranking, and Link-the-Wiki Tracks	331
<i>Kelly Y. Itakura and Charles L.A. Clarke</i>	
A Machine Learning Approach to Link Prediction for Interlinked Documents	342
<i>Milly Kc, Rowena Chau, Markus Hagenbuchner, Ah Chung Tsoi, and Vincent Lee</i>	

Question Answering Track

Overview of the 2009 QA Track: Towards a Common Task for QA, Focused IR and Automatic Summarization Systems	355
<i>Veronique Moriceau, Eric SanJuan, Xavier Tannier, and Patrice Bellot</i>	

XML Mining Track

Overview of the INEX 2009 XML Mining Track: Clustering and Classification of XML Documents	366
<i>Richi Nayak, Christopher M. De Vries, Sangeetha Kutty, Shlomo Geva, Ludovic Denoyer, and Patrick Gallinari</i>	
Exploiting Index Pruning Methods for Clustering XML Collections	379
<i>Ismail Sengor Altingovde, Duygu Atilgan, and Özgür Ulusoy</i>	
Multi-label Wikipedia Classification with Textual and Link Features	387
<i>Boris Chidlovskii</i>	
Link-Based Text Classification Using Bayesian Networks	397
<i>Luis M. de Campos, Juan M. Fernández-Luna, Juan F. Huete, Andrés R. Masegosa, and Alfonso E. Romero</i>	
Clustering with Random Indexing K-tree and XML Structure	407
<i>Christopher M. De Vries, Shlomo Geva, and Lance De Vine</i>	
Utilising Semantic Tags in XML Clustering	416
<i>Sangeetha Kutty, Richi Nayak, and Yuefeng Li</i>	
UJM at INEX 2009 XML Mining Track	426
<i>Christine Largeron, Christophe Moulin, and Mathias Géry</i>	
BUAP: Performance of K-Star at the INEX'09 Clustering Task	434
<i>David Pinto, Mireya Tovar, Darnes Vilarino, Beatriz Beltrán, Héctor Jiménez-Salazar, and Basilia Campos</i>	
Extended VSM for XML Document Classification Using Frequent Subtrees	441
<i>Jianwu Yang and Songlin Wang</i>	
Supervised Encoding of Graph-of-Graphs for Classification and Regression Problems	449
<i>Shu Jia Zhang, Markus Hagenbuchner, Franco Scarselli, and Ah Chung Tsoi</i>	

Erratum

Overview of the INEX 2009 Ad Hoc Track	E1
<i>Shlomo Geva, Jaap Kamps, Miro Lethonen, Ralf Schenkel, James A. Thom, and Andrew Trotman</i>	

Author Index	463
-------------------------------	-----