Lecture Notes in Computer Science 6203

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

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

Springer

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

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 particants, 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 Pharo	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

Ad Hoc Track

Book Track

Efficiency Track

Entity Ranking Track

Interactive Track

Link the Wiki Track

Question Answering Track

XML Mining Track

Erratum