Lecture Notes in Computer Science 5631

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

Shlomo Geva Jaap Kamps
Andrew Trotman (Eds.)

Advances in Focused Retrieval

7th International Workshop of the Initiative
for the Evaluation of XML Retrieval, INEX 2008
Dagstuhl Castle, Germany, December 15-18, 2008
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: 2009933190

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 3-642-03760-7 Springer Berlin Heidelberg New York
ISBN-13 978-3-642-03760-3 Springer Berlin Heidelberg New York

Typesetting: Camera-ready by author, data conversion by Scientific Publishing Services, Chennai, India
Printed on acid-free paper SPIN: 12727117 06/3180 5 4 3 2 1 0

Foreword

I write with pleasure this foreword to the proceedings of the 7th workshop of the Initiative for the Evaluation of XML Retrieval (INEX). The increased adoption of XML as the standard for representing a document structure has led to the development of retrieval systems that are aimed at effectively accessing XML documents. Providing effective access to large collections of XML documents is therefore a key issue for the success of these systems. INEX aims to provide the necessary methodological means and worldwide infrastructures for evaluating how good XML retrieval systems are.

Since its launch in 2002, INEX has grown both in terms of number of participants and its coverage of the investigated retrieval tasks and scenarios. In 2002, INEX started with 49 registered participating organizations, whereas this number was more than 100 for 2008. In 2002, there was one main track, concerned with the ad hoc retrieval task, whereas in 2008, seven tracks in addition to the main ad hoc track were investigated, looking at various aspects of XML retrieval, from book search to entity ranking, including interaction aspects.

INEX follows the predominant approach in information retrieval of evaluating retrieval approaches using a test collection constructed specifically for that purpose and associated effectiveness measures. A test collection usually consists of a set of documents, user requests (topics), and relevance assessments, which specify the set of "right answers" for the requests. Throughout the years, INEX faced a range of challenges regarding the evaluation of XML retrieval systems, as the consideration of the structure led to many complex issues, which were not always identified at the beginning (e.g., the issue of overlap and a counting-based effectiveness measure, the difficulty in consistently assessing elements using a four-graded two-dimensional scale). In addition, limited funding was available, for example, to pay assessors. This led to a research problem in itself, namely, how to elicit quality assessments in order for the test collections to be reusable. As a result, different theories and methods for the evaluation of XML retrieval were developed and tested at INEX, e.g., the definition of relevance, a plethora of effectiveness measures, leading to a now-stable evaluation setup and a rich history of learned lessons. This is now allowing researchers worldwide to make further progress in XML retrieval, including the investigation of other research questions in XML retrieval, for instance, efficiency in 2008.

What I have greatly enjoyed with INEX is the people working together not only to develop approaches for XML retrieval, but also methodologies for evaluating XML retrieval. Many of the people that actually joined INEX to test their XML retrieval approaches got *hooked* on the problem of properly evaluating XML retrieval approaches. Three of them are the current INEX organizers, Shlomo Geva, Jaap Kamps, and Andrew Trotman. I am very glad that they got hooked, as they are dedicated and, even more importantly, enthusiastic people

with extensive expertise in both building XML retrieval systems and evaluating them. INEX is in very safe hands with them. Looking at the current proceedings, I am delighted to see so many excellent works from people from all over the world. Many of them met at the annual workshop, and I heard it was a great success. Well done to all.

April 2009 Mounia Lalmas

Preface

Welcome to the 7th workshop of the Initiative for the Evaluation of XML Retrieval (INEX)! Now, in its seventh year, INEX is one of the established evaluation forums in information retrieval (IR), with 150 organizations worldwide registering and over 50 groups participating actively in the different 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.

Information on the Web is a mixture of text, multimedia, and metadata, with a clear internal structure, usually formatted according to the eXtensible Markup Language (XML) standard, or another related W3C standard. 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 2008 was an exciting year for INEX, and brought a lot of changes. Seven research tracks were included, which studied different aspects of focused information access: Ad Hoc, Book, Efficiency, Entity Ranking, Interactive (iTrack), Link the Wiki, and XML Mining. The aim of the INEX 2008 workshop was to bring together researchers who participated in the INEX 2008 campaign. During the past year, participating organizations contributed to the building of a large-scale XML test collection by creating topics, performing retrieval runs, and providing relevance assessments. The workshop concluded the results of this large-scale 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 2008. We accepted a total of 49 out of 53 papers, yielding a 92% acceptance rate.

This was also the seventh INEX Workshop to be held at the *Schloss Dagstuhl – Leibniz Center for Informatics*, providing a unique setting where informal interaction and discussion occurs naturally and frequently. This has been essential to the growth of INEX over the years, and we feel honored and privileged that *Dagstuhl* housed the INEX 2008 Workshop. Finally, INEX was 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 2009

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	University of Glasgow, 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	General, Queensland University of Technology, Australia
Jaap Kamps	General, University of Amsterdam, The Netherlands
Andrew Trotman	General, University of Otago, New Zealand
Ludovic Denoyer	Document Collection, University Paris 6, France
Ralf Schenkel	Document Exploration, Max-Planck-Institut für Informatik, Germany
Martin Theobald	Document Exploration, Stanford University, USA

Book

Antoine Doucet	University of Caen, France
Gabriella Kazai	Microsoft Research Limited, Cambridge, UK
Monica Landoni	University of Strathclyde, UK

Efficiency

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

Entity Ranking

Gianluca Demartini	L3S, Leibniz Universität Hannover, Germany
Tereza Iofciu	L3S, Leibniz Universität Hannover, Germany
Arjen de Vries	CWI, The Netherlands
Jianhan Zhu	University College London, UK

Interactive (iTrack)

Nisa Fachry	University of Amsterdam, The Netherlands
Ragnar Nordlie	Oslo University College, Norway
Nils Pharo	Oslo University College, Norway

Link the Wiki

Shlomo Geva	Queensland University of Technology, Australia
Wei-Che (Darren) Huang	Queensland University of Technology, Australia
Andrew Trotman	University of Otago, New Zealand

XML Mining

Ludovic Denoyer	University Paris 6, France
Patrick Gallinari	University Paris 6, France

Table of Contents

Ad Hoc Track

Book Track

Efficiency Track

Entity Ranking Track

Interactive Track

Link the Wiki Track

XML Mining Track