Lecture Notes in Computer Science 4862

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

Norbert Fuhr Jaap Kamps Mounia Lalmas
Andrew Trotman (Eds.)

Focused Access to XML Documents

6th International Workshop of the Initiative
for the Evaluation of XML Retrieval, INEX 2007
Dagstuhl Castle, Germany, December 17-19, 2007
Revised and Selected Papers

Springer

Volume Editors

Norbert Fuhr
University of Duisburg-Essen, Department of Computational
and Cognitive Sciences, 47048 Duisburg, Germany
E-mail: norbert.fuhr@uni-due.de

Jaap Kamps
University of Amsterdam, Archives and Information Studies/Humanities
1012 Amsterdam, The Netherlands
E-mail: kamps@science.uva.nl

Mounia Lalmas
University of London, Department of Computer Science
Queen Mary, E1 4NS, London, United Kingdom
E-mail: mounia@dcs.qmul.ac.uk

Andrew Trotman
University of Otago, Department of Computer Science
9015 Dunedin, New Zealand
E-mail: andrew@cs.otago.ac.nz

Library of Congress Control Number: 2008934300

CR Subject Classification (1998): H.3, H.4, H.2

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

ISSN 0302-9743
ISBN-10 3-540-85901-2 Springer Berlin Heidelberg New York
ISBN-13 978-3-540-85901-7 Springer Berlin Heidelberg New York

Typesetting: Camera-ready by author, data conversion by Scientific Publishing Services, Chennai, India
Printed on acid-free paper SPIN: 12522002 06/3180 5 4 3 2 1 0

Preface

Welcome to the proceedings of the 6th workshop of the Initiative for the Evaluation of XML Retrieval (INEX)! Now in its sixth year, INEX has become an established evaluation forum for XML information retrieval (IR), with over 100 participating organizations worldwide. Its aim is to provide an infrastructure, in the form of a large XML test collection and appropriate scoring methods, for the evaluation of XML IR systems.

XML IR is playing an increasingly important role in many information access systems (e.g., digital libraries, web, intranet) where content is becoming more and more 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.

The aim of the INEX 2007 workshop was to bring together researchers in the field of XML IR who participated in the INEX 2007 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 brought together the results of this large-scale effort, summarized and addressed the issues encountered, and devised a work plan for the future evaluation of XML retrieval systems.

In total seven research tracks were included in INEX 2007. These studied different aspects of XML information access: ad-hoc, document mining, multimedia, heterogeneous, entity ranking, book search, and link-the-wiki. The consolidation of the existing tracks, and the expansion to new areas offered by the new tracks has enabled INEX to extend its scope. This volume contains 37 papers selected from 50 submitted ones (74% acceptance rate). Each paper was peer-reviewed.

INEX is funded by the DELOS Network of Excellence on Digital Libraries, to which we are very thankful. We thank Schloss Dagstuhl (Leibniz Center for Informatics) for housing the INEX workshop in its unique and inspiring atmosphere, and Springer for publishing the results of INEX 2007 in these final proceedings. We gratefully thank the organizers of the various tasks and tracks who did a superb job. Finally, special thanks go to the participating organizations and people for their contribution.

May 2008

Norbert Fuhr
Jaap Kamps
Mounia Lalmas
Andrew Trotman

Organization

Project Leaders

Norbert Fuhr	University of Duisburg-Essen, Germany
Mounia Lalmas	Queen Mary, University of London, UK
Andrew Trotman	University of Otago, New Zealand

Contact People

Saadia Malik	University of Duisburg-Essen, Germany
Zoltán Szlávik	Queen Mary, University of London, UK

Wikipedia Document Collection

Ludovic Denoyer	University Paris 6, France

Document Exploration

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

Topic Format Specification

Birger Larsen	Royal School of Library and Information Science, Denmark
Andrew Trotman	University of Otago, New Zealand

Task Description

Jaap Kamps	University of Amsterdam, The Netherlands
Charles Clarke	University of Waterloo, Canada

Online Relevance Assessment Tool

Benjamin Piwowarski	Yahoo! Research Latin America, Chile

Effectiveness Measures

Gabriella Kazai	Microsoft Research Cambridge, UK
Benjamin Piwowarski	Yahoo! Research Latin America, Chile

Jaap Kamps	University of Amsterdam, The Netherlands
Jovan Pehcevski	INRIA Rocquencourt, France
Stephen Robertson	Microsoft Research Cambridge, UK

Document Mining Track

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

Multimedia Track

Thijs Westerveld	CWI, The Netherlands
Theodora Tsikrika	CWI, The Netherlands

Entity Ranking Track

Arjen de Vries	CWI, The Netherlands
Nick Craswell	Microsoft Research Cambridge, UK
Mounia Lalmas	Queen Mary, University of London, UK
James A. Thom	RMIT University, Australia
Anne-Marie Vercoustre	INRIA Rocquencourt, France

Link the Wiki Track

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

Book Search

Gabriella Kazai	Microsoft Research Cambridge, UK
Antoine Doucet	University of Caen, France

Table of Contents

Ad Hoc Track

Book Search Track

XML-Mining Track

Entity Ranking Track

Interactive Track

Link-the-Wiki Track

Multimedia Track