Lecture Notes in Computer Science 3493

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

Norbert Fuhr   Mounia Lalmas
Saadia Malik   Zoltán Szlávik (Eds.)

Advances in XML Information Retrieval

Third International Workshop of the Initiative
for the Evaluation of XML Retrieval, INEX 2004
Dagstuhl Castle, Germany, December 6-8, 2004
Revised Selected Papers

Springer

Volume Editors

Norbert Fuhr
University of Duisburg-Essen
Faculty of Engineering Sciences, Information Systems
47048 Duisburg, Germany
E-mail: fuhr@uni-duisburg.de

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

Saadia Malik
University of Duisburg-Essen
Faculty of Engineering Sciences, Information Systems
47048 Duisburg, Germany
E-mail: malik@is.informatik.uni-duisburg.de

Zoltán Szlávik
Queen Mary University of London, Department of Computer Science
London E1 4NS, England, United Kingdom
E-mail: zolley@dcs.qmul.ac.uk

Library of Congress Control Number: 2005926830

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

ISSN 0302-9743
ISBN-10 3-540-26166-4 Springer Berlin Heidelberg New York
ISBN-13 978-3-540-26166-7 Springer Berlin Heidelberg New York

Springer is a part of Springer Science+Business Media

springeronline.com

Typesetting: Camera-ready by author, data conversion by Scientific Publishing Services, Chennai, India
Printed on acid-free paper SPIN: 11424550 06/3142 5 4 3 2 1 0

Preface

The ultimate goal of many information access systems (e.g., digital libraries, the Web, intranets) is to provide the right content to their end-users. This content is increasingly a mixture of text, multimedia, and metadata, and is formatted according to the adopted –W3C standard for information repositories, the so-called eXtensible Markup Language (XML). Whereas 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 campaign (Initiative for the Evaluation of XML Retrieval), which was set up at the beginning of 2002, is to establish infrastructures, XML test suites, and appropriate measurements for evaluating the performance of information retrieval systems that aim at giving effective access to XML content. More precisely, the goal of the INEX initiative is to provide means, in the form of a large XML test collection and appropriate scoring methods, for the evaluation of content-oriented XML retrieval systems.

INEX 2004 was responsible for a range of evaluation activities in the field of XML information retrieval, with five tracks: (1) *Ad Hoc Retrieval Track*, the main track, which can be regarded as a simulation of how a digital library might be used, where a static set of XML documents and their components is searched using a new set of queries (topics) containing both content and structural conditions; (2) *Interactive Track*, which aimed to investigate the behavior of users when interacting with components of XML documents; (3) *Heterogeneous Collection Track*, where retrieval is based on a collection comprising various XML subcollections from different digital libraries, as well as material from other resources; (4) *Relevance Feedback Track*, dealing with relevance feedback methods for XML; and (5) *Natural Language Track*, where natural language formulations of structural conditions of queries have to be answered.

The INEX 2004 workshop, held at Schloss Dagstuhl (Germany), 6–8 December 2004, brought together researchers in the field of XML retrieval who participated in the INEX 2004 evaluation campaign. Participants were able to present and discuss their approaches to XML retrieval. These proceedings contain revised papers describing work carried out during INEX 2004 in the various tracks by the participants.

INEX is partly funded by the DELOS Network of Excellence on Digital Libraries, to which we are very thankful. We would also like to thank the IEEE Computer Society for providing us the XML document collection. Special thanks go to Shlomo Geva for setting up the WIKI server and Gabriella Kazai for helping with the various documentation. We gratefully acknowledge the involvement of Börkur Sigurbjörnsson and Andrew Trotman (topic format specification), Benjamin Piwowarski (online

assessment tool), and Gabriella Kazai and Arjen de Vries (metrics). The organizers of the various tracks did a great job and their work is greatly appreciated: Anastasios Tombros, Birger Larsen, Thomas Rölleke, Carolyn Crouch, Shlomo Geva and Tony Sahama. Finally, we would like to thank the participating organizations and people for their participation in INEX 2004.

March 2005

Norbert Fuhr
Mounia Lalmas
Saadia Malik
Zoltán Szlávik

Organizers

Organizers

Project Leaders

Norbert Fuhr, University of Duisburg-Essen
Mounia Lalmas, Queen Mary University of London

Contact Person

Saadia Malik, University of Duisburg-Essen

Topic Format Specification

Börkur Sigurbjörnsson, University of Amsterdam
Andrew Trotman, University of Otago

Online Relevance Assessment Tool

Benjamin Piwowarski, University of Chile

Metrics

Gabriella Kazai, Queen Mary University of London
Arjen P. de Vries, Centre for Mathematics and Computer Science

Interactive Track

Birger Larsen, Royal School of Library and Information Science
Saadia Malik, University of Duisburg-Essen
Anastasios Tombros, Queen Mary University of London

Relevance Feedback Track

Carolyn Crouch, University of Minnesota-Duluth
Mounia Lalmas, Queen Mary University of London

Heterogeneous Collection Track

Thomas Rölleke, Queen Mary University of London
Zoltán Szlávik, Queen Mary University of London

Natural Language Processing

Shlomo Geva, Queensland University of Technology
Tony Sahama, Queensland University of Technology

Table of Contents

Methodology

Ad Hoc Retrieval

Ad Hoc Retrieval and Relevance Feedback

Relevance Feedback

Ad Hoc Retrieval and Heterogeneous Document Collection

Heterogeneous Document Collection

Natural Language Processing of Topics

Interactive Studies