

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

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

University of Dortmund, Germany

Madhu Sudan

Massachusetts Institute of Technology, MA, USA

Demetri Terzopoulos

New York University, NY, USA

Doug Tygar

University of California, Berkeley, CA, USA

Moshe Y. Vardi

Rice University, Houston, TX, USA

Gerhard Weikum

Max-Planck Institute of Computer Science, Saarbruecken, Germany

Norbert Fuhr Mounia Lalmas
Saadia Malik Zoltán Szilávik (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 Szilávik
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

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 is a part of Springer Science+Business Media

springeronline.com

© Springer-Verlag Berlin Heidelberg 2005
Printed in Germany

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 Szilávik, Queen Mary University of London

Natural Language Processing

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

Table of Contents

Overview of INEX 2004 <i>Saadia Malik, Mounia Lalmas, Norbert Fuhr</i>	1
Methodology	
Narrowed Extended XPath I (NEXI) <i>Andrew Trotman, Börkur Sigurbjörnsson</i>	16
NEXI, Now and Next <i>Andrew Trotman, Börkur Sigurbjörnsson</i>	41
If INEX Is the Answer, What Is the Question? <i>Richard A. O'Keefe</i>	54
Reliability Tests for the XCG and inex-2002 Metrics <i>Gabriella Kazai, Mounia Lalmas, Arjen de Vries</i>	60
Ad Hoc Retrieval	
Component Ranking and Automatic Query Refinement for XML Retrieval <i>Yosi Mass, Matan Mandelbrod</i>	73
MultiText Experiments for INEX 2004 <i>Charles L.A. Clarke, Philip L. Tilker</i>	85
Logic-Based XML Information Retrieval for Determining the Best Element to Retrieve <i>Maryam Karimzadegan, Jafar Habibi, Farhad Oroumchian</i>	88
An Algebra for Structured Queries in Bayesian Networks <i>Jean-Noël Vittaut, Benjamin Piwowarski, Patrick Gallinari</i>	100
IR of XML Documents – A Collective Ranking Strategy <i>Maha Salem, Alan Woodley, Shlomo Geva</i>	113
TRIX 2004 – Struggling with the Overlap <i>Jaana Kekäläinen, Marko Junkkari, Paavo Arvola, Timo Aalto</i>	127
The Utrecht Blend: Basic Ingredients for an XML Retrieval System <i>Roelof van Zwol, Frans Wiering, Virginia Dignum</i>	140

Hybrid XML Retrieval Revisited <i>Jovan Pehcevski, James A. Thom, S.M.M. Tahaghoghi, Anne-Marie Vercoustre</i>	153
Analyzing the Properties of XML Fragments Decomposed from the INEX Document Collection <i>Kenji Hatano, Hiroko Kinutani, Toshiyuki Amagasa, Yasuhiro Mori, Masatoshi Yoshikawa, Shunsuke Uemura</i>	168
A Voting Method for XML Retrieval <i>Gilles Hubert</i>	183
Mixture Models, Overlap, and Structural Hints in XML Element Retrieval <i>Börkur Sigurbjörnsson, Jaap Kamps, Maarten de Rijke</i>	196
GPX – Gardens Point XML Information Retrieval at INEX 2004 <i>Shlomo Geva</i>	211
Hierarchical Language Models for XML Component Retrieval <i>Paul Ogilvie, Jamie Callan</i>	224
Ranked Retrieval of Structured Documents with the S-Term Vector Space Model <i>Felix Weigel, Klaus U. Schulz, Holger Meuss</i>	238
Merging XML Indices <i>Gianni Amati, Claudio Carpineto, Giovanni Romano</i>	253
DocBase – The INEX Evaluation Experience <i>Sriram Mohan, Arijit Sengupta</i>	261
Ad Hoc Retrieval and Relevance Feedback	
TIJAH at INEX 2004 Modeling Phrases and Relevance Feedback <i>Vojkan Mihajlović, Georgina Ramírez, Arjen P. de Vries, Djoerd Hiemstra, Henk Ernst Blok</i>	276
Flexible Retrieval Based on the Vector Space Model <i>Carolyn J. Crouch, Aniruddha Mahajan, Archana Bellamkonda</i>	292
Relevance Feedback	
Relevance Feedback for XML Retrieval <i>Yosi Mass, Matan Mandelbrod</i>	303

Ad Hoc Retrieval and Heterogeneous Document Collection

A Universal Model for XML Information Retrieval <i>Maria Izabel M. Azevedo, Lucas Pantuza Amorim, Nívio Ziviani</i>	311
Cheshire II at INEX '04: Fusion and Feedback for the Adhoc and Heterogeneous Tracks <i>Ray R. Larson</i>	322
Using a Relevance Propagation Method for Adhoc and Heterogeneous Tracks at INEX 2004 <i>Karen Sauvagnat, Mohand Boughanem</i>	337

Heterogeneous Document Collection

Building and Experimenting with a Heterogeneous Collection <i>Zoltán Szlávik, Thomas Rölleke</i>	349
A Test Platform for the INEX Heterogeneous Track <i>Serge Abiteboul, Ioana Manolescu, Benjamin Nguyen, Nicoleta Preda</i>	358
EXTIRP 2004: Towards Heterogeneity <i>Miro Lehtonen</i>	372

Natural Language Processing of Topics

NLPX at INEX 2004 <i>Alan Woodley, Shlomo Geva</i>	382
Analysing Natural Language Queries at INEX 2004 <i>Xavier Tannier, Jean-Jacques Girardot, Mihaela Mathieu</i>	395

Interactive Studies

The Interactive Track at INEX 2004 <i>Anastasios Tombros, Birger Larsen, Saadia Malik</i>	410
Interactive Searching Behavior with Structured XML Documents <i>Heesop Kim, Heejung Son</i>	424
Author Index	437