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
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

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

Overview of INEX 2004  
*Saadia Malik, Mounia Lalmas, Norbert Fuhr* ................................................................. 1

## Methodology

Narrowed Extended XPath I (NEXI)  
*Andrew Trotman, Börkur Sigurbjörnsson* ........................................................................ 16

NEXI, Now and Next  
*Andrew Trotman, Börkur Sigurbjörnsson* ........................................................................ 41

If INEX Is the Answer, What Is the Question?  
*Richard A. O'Keefe* ........................................................................................................... 54

Reliability Tests for the XCG and inex-2002 Metrics  
*Gabriella Kazai, Mounia Lalmas, Arjen de Vries* ............................................................ 60

## Ad Hoc Retrieval

Component Ranking and Automatic Query Refinement for XML Retrieval  
*Yosi Mass, Matan Mandelbrod* ........................................................................................... 73

MultiText Experiments for INEX 2004  
*Charles L.A. Clarke, Philip L. Tilker* .................................................................................. 85

Logic-Based XML Information Retrieval for Determining the Best Element to Retrieve  
*Maryam Karimzadegan, Jafar Habibi, Farhad Oroumchian* ........................................... 88

An Algebra for Structured Queries in Bayesian Networks  
*Jean-Noël Vittaut, Benjamin Piwowarski, Patrick Gallinari* .............................................. 100

IR of XML Documents – A Collective Ranking Strategy  
*Maha Salem, Alan Woodley, Shlomo Geva* ....................................................................... 113

TRIX 2004 – Struggling with the Overlap  
*Jaana Kekäläinen, Marko Junkkari, Paavo Arvola, Timo Aalto* ......................................... 127

The Utrecht Blend: Basic Ingredients for an XML Retrieval System  
*Roelof van Zwol, Frans Wiering, Virginia Dignum* ......................................................... 140
<table>
<thead>
<tr>
<th>Title</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>Hybrid XML Retrieval Revisited</td>
<td>153</td>
</tr>
<tr>
<td>Jovan Pehcevski, James A. Thom, S.M.M. Tahaghoghi, Anne-Marie Vercoustre</td>
<td></td>
</tr>
<tr>
<td>Analyzing the Properties of XML Fragments Decomposed from the INEX Document Collection</td>
<td>168</td>
</tr>
<tr>
<td>Kenji Hatano, Hiroko Kinutani, Toshiyuki Amagasa, Yasuhiro Mori, Masatoshi Yoshikawa, Shunsuke Uemura</td>
<td></td>
</tr>
<tr>
<td>A Voting Method for XML Retrieval</td>
<td>183</td>
</tr>
<tr>
<td>Gilles Hubert</td>
<td></td>
</tr>
<tr>
<td>Mixture Models, Overlap, and Structural Hints in XML Element Retrieval</td>
<td>196</td>
</tr>
<tr>
<td>Börkur Sigurbjörnsson, Jaap Kamps, Maarten de Rijke</td>
<td></td>
</tr>
<tr>
<td>GPX – Gardens Point XML Information Retrieval at INEX 2004</td>
<td>211</td>
</tr>
<tr>
<td>Shlomo Geva</td>
<td></td>
</tr>
<tr>
<td>Hierarchical Language Models for XML Component Retrieval</td>
<td>224</td>
</tr>
<tr>
<td>Paul Ogilvie, Jamie Callan</td>
<td></td>
</tr>
<tr>
<td>Ranked Retrieval of Structured Documents with the S-Term Vector Space Model</td>
<td>238</td>
</tr>
<tr>
<td>Felix Weigel, Klaus U. Schulz, Holger Meuss</td>
<td></td>
</tr>
<tr>
<td>Merging XML Indices</td>
<td>253</td>
</tr>
<tr>
<td>Gianni Amati, Claudio Carpineto, Giovanni Romano</td>
<td></td>
</tr>
<tr>
<td>DocBase – The INEX Evaluation Experience</td>
<td>261</td>
</tr>
<tr>
<td>Sriram Mohan, Arijit Sengupta</td>
<td></td>
</tr>
<tr>
<td>Ad Hoc Retrieval and Relevance Feedback</td>
<td>276</td>
</tr>
<tr>
<td>TIJAH at INEX 2004 Modeling Phrases and Relevance Feedback</td>
<td></td>
</tr>
<tr>
<td>Vojkan Mihajlović, Georgina Ramírez, Arjen P. de Vries, Djoerd Hiemstra, Henk Ernst Blok</td>
<td></td>
</tr>
<tr>
<td>Flexible Retrieval Based on the Vector Space Model</td>
<td>292</td>
</tr>
<tr>
<td>Carolyn J. Crouch, Aniruddha Mahajan, Archana Bellamkonda</td>
<td></td>
</tr>
<tr>
<td>Relevance Feedback</td>
<td>303</td>
</tr>
<tr>
<td>Relevance Feedback for XML Retrieval</td>
<td></td>
</tr>
<tr>
<td>Yosi Mass, Matan Mandelbrod</td>
<td></td>
</tr>
</tbody>
</table>
Ad Hoc Retrieval and Heterogeneous Document Collection

A Universal Model for XML Information Retrieval
Maria Izabel M. Azevedo, Lucas Pantuza Amorim, Nívio Ziviani......................... 311

Cheshire II at INEX ’04: Fusion and Feedback for the Adhoc and Heterogeneous Tracks
Ray R. Larson........................................................................................................ 322

Using a Relevance Propagation Method for Adhoc and Heterogeneous Tracks at INEX 2004
Karen Sauvagnat, Mohand Boughanem................................................................. 337

Heterogeneous Document Collection

Building and Experimenting with a Heterogeneous Collection
Zoltán Szlávik, Thomas Rölleke............................................................................. 349

A Test Platform for the INEX Heterogeneous Track
Serge Abiteboul, Ioana Manolescu, Benjamin Nguyen, Nicoleta Preda............... 358

EXTIRP 2004: Towards Heterogeneity
Miro Lehtonen........................................................................................................ 372

Natural Language Processing of Topics

NLPX at INEX 2004
Alan Woodley, Shlomo Geva............................................................................... 382

Analysing Natural Language Queries at INEX 2004
Xavier Tannier, Jean-Jacques Girardot, Mihaela Mathieu................................. 395

Interactive Studies

The Interactive Track at INEX 2004
Anastasios Tombros, Birger Larsen, Saadia Malik............................................ 410

Interactive Searching Behavior with Structured XML Documents
Heesop Kim, Heejung Son..................................................................................... 424

Author Index.......................................................................................................... 437