Lecture Notes in Computer Science

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
   University of California, Los Angeles, CA, 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  Gabriella Kazai (Eds.)

Advances in
XML Information
Retrieval and Evaluation

4th International Workshop of the Initiative
for the Evaluation of XML Retrieval, INEX 2005
Dagstuhl Castle, Germany, November 28-30, 2005
Revised Selected Papers
Preface

Content-oriented XML retrieval has been receiving increasing interest due to the widespread use of eXtensible Markup Language (XML), which is becoming a standard document format on the Web, in digital libraries, and publishing. By exploiting the enriched source of syntactic and semantic information that XML markup provides, XML information retrieval (IR) systems aim to implement a more focused retrieval strategy and return document components, so-called XML elements – instead of complete documents – in response to a user query. This focused retrieval approach is of particular benefit for collections containing long documents or documents covering a wide variety of topics (e.g., books, user manuals, legal documents, etc.), where users’ effort to locate relevant content can be reduced by directing them to the most relevant parts of the documents. Implementing this, more focused, retrieval paradigm means that an XML IR system needs not only to find relevant information in the XML documents, but it also has to determine the appropriate level of granularity to be returned to the user. In addition, the relevance of a retrieved component may be dependent on meeting both content and structural query conditions.

Evaluating the effectiveness of XML retrieval systems, hence, requires a test collection where relevance assessments are provided according to a relevance criterion, which takes into account the imposed structural aspects. In 2002, the INitiative for the Evaluation of XML Retrieval (INEX) started to address these issues. The aim of the INEX initiative is to establish an infrastructure and provide means, in the form of a large XML test collection and appropriate scoring methods, for the evaluation of content-oriented XML retrieval systems. Now, in its fourth year, INEX is an established evaluation forum for XML IR, with over 50 participating organizations worldwide.

2005 was an exciting year for INEX, and brought with it many changes and new aspects to the evaluation. Several new tracks and tasks, a new relevance assessment procedure and new evaluation measures were introduced. In total, seven research tracks were included in INEX 2005, which studied different aspects of XML information access: Ad-hoc retrieval, Interactive, Relevance Feedback, Heterogeneous, Natural Language Processing, and two new tracks for 2005, Multimedia and Document Mining.

The INEX 2005 workshop, held at Schloss Dagstuhl (Germany), November 28–30, 2005, brought together researchers in the field of XML retrieval, who participated in the INEX 2005 evaluation campaign. Participants were able to present and discuss their approach to XML retrieval and evaluation. These proceedings contain revised papers describing work carried out during INEX 2005 in the various tracks.

INEX is 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 and the Lonely Planet for providing us the data sets, which were used to build two of the XML collections used in INEX 2005. We gratefully thank the organizers of the various tasks and tracks who did a superb job, their work is greatly appreciated. Finally, special thanks go to the participating organizations and people for their contributions.

March 2005

Norbert Fuhr
Mounia Lalmas
Saadia Malik
Gabriella Kazai
Organization

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

Contact Persons
Saadia Malik  University of Duisburg-Essen
Zoltan Szlavik  Queen Mary, University of London

Topic Format Specification
Börkur Sigurbjörnsson  University of Amsterdam
Andrew Trotman  University of Otago

Online Relevance Assessment Tool
Benjamin Piwowarski  Universidad de Chile

Evaluation Measures
Gabriella Kazai  Queen Mary, University of London
Arjen P. de Vries  CWI
Paul Ogilvie  Carnegie Mellon University
Benjamin Piwowarski  Universidad de Chile

Relevance Feedback Task
Yosi Mass  IBM Research Lab
Carolyn J. Crouch  University of Minnesota Duluth

Natural Language Processing Task
Shlomo Geva  Queensland University of Technology
Alan Woodley  Queensland University of Technology

Heterogeneous Collection Track
Ray Larson  University of California, Berkeley
Interactive Track

Birger Larsen  Royal School of LIS
Anastasios Tombros  Queen Mary, University of London
Saadia Malik  University of Duisburg-Essen

Document Mining Track

Ludovic Denoyer  Laboratoire d’Informatique de Paris 6
Patrick Gallinari  Laboratoire d’Informatique de Paris 6
Anne-Marie Vercoustre  INRIA

XML Multimedia Track

Roelof van Zwol  Utrecht University
Gabriella Kazai  Queen Mary, University of London
Mounia Lalmas  Queen Mary, University of London
Table of Contents

Methodology

Overview of INEX 2005
   Saadia Malik, Gabriella Kazai, Mounia Lalmas, Norbert Fuhr ....... 1

INEX 2005 Evaluation Measures
   Gabriella Kazai, Mounia Lalmas ........................................ 16

EPRUM Metrics and INEX 2005
   Benjamin Piwowarski .................................................. 30

HiXEval: Highlighting XML Retrieval Evaluation
   Jovan Pehcevski, James A. Thom .................................... 43

The Interpretation of CAS
   Andrew Trotman, Mounia Lalmas ..................................... 58

Multiple Tracks

TIJAH Scratches INEX 2005: Vague Element Selection, Image Search, Overlap, and Relevance Feedback
   Vojkan Mihajlović, Georgina Ramírez, Thijs Westerveld,
   Djoerd Hienstra, Henk Ernst Blok, Arjen P. de Vries .............. 72

XFIRM at INEX 2005: Ad-Hoc and Relevance Feedback Tracks
   Karen Sauvagnat, Lobna Hlaoua, Mohand Boughanem ............. 88

Ad-Hoc Track

The Effect of Structured Queries and Selective Indexing on XML Retrieval
   Börkur Sigurbjörnsson, Jaap Kamps ................................. 104

Searching XML Documents – Preliminary Work
   Marcus Hassler, Abdelhamid Bouchachia ............................ 119

Query Evaluation with Structural Indices
   Paavo Arvola, Jaana Kekäläinen, Marko Junkkari .................... 134
<table>
<thead>
<tr>
<th>Title</th>
<th>Author(s)</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>B³-SDR and Effective Use of Structural Hints</td>
<td>Roelof van Zwol</td>
<td>146</td>
</tr>
<tr>
<td>Field-Weighted XML Retrieval Based on BM25</td>
<td>Wei Lu, Stephen Robertson, Andrew MacFarlane</td>
<td>161</td>
</tr>
<tr>
<td>XML Retrieval Based on Direct Contribution of Query Components</td>
<td>Gilles Hubert</td>
<td>172</td>
</tr>
<tr>
<td>Using the INEX Environment as a Test Bed for Various User Models for XML Retrieval</td>
<td>Yosi Mass, Matan Mandelbrod</td>
<td>187</td>
</tr>
<tr>
<td>The University of Kaiserslautern at INEX 2005</td>
<td>Philipp Dopichaj</td>
<td>196</td>
</tr>
<tr>
<td>Parameter Estimation for a Simple Hierarchical Generative Model for XML Retrieval</td>
<td>Paul Ogilvie, Jamie Callan</td>
<td>211</td>
</tr>
<tr>
<td>Probabilistic Retrieval, Component Fusion and Blind Feedback for XML Retrieval</td>
<td>Ray R. Larson</td>
<td>225</td>
</tr>
<tr>
<td>GPX – Gardens Point XML IR at INEX 2005</td>
<td>Shlomo Geva</td>
<td>240</td>
</tr>
<tr>
<td>Implementation of a High-Speed and High-Precision XML Information Retrieval System on Relational Databases</td>
<td>Kei Fujimoto, Toshiyuki Shimizu, Norimasa Terada, Kenji Hatano, Yu Suzuki, Toshiyuki Amagasa, Hiroko Kinutani, Masatoshi Yoshikawa</td>
<td>254</td>
</tr>
<tr>
<td>The Dynamic Retrieval of XML Elements</td>
<td>Carolyn J. Crouch, Sudip Khanna, Poorva Potnis, Nagendra Doddapaneni</td>
<td>268</td>
</tr>
<tr>
<td>TopX and XXL at INEX 2005</td>
<td>Martin Theobald, Ralf Schenkel, Gerhard Weikum</td>
<td>282</td>
</tr>
<tr>
<td>When a Few Highly Relevant Answers Are Enough</td>
<td>Miro Lehtonen</td>
<td>296</td>
</tr>
<tr>
<td>RMIT University at INEX 2005: Ad Hoc Track</td>
<td>Jovan Pehcevski, James A. Thom, S.M.M. Tahaghoghi</td>
<td>306</td>
</tr>
</tbody>
</table>
SIRIUS: A Lightweight XML Indexing and Approximate Search System at INEX 2005  
Eugen Popovici, Gildas Ménier, Pierre-François Marteau ............... 321

Machine Learning Ranking and INEX’05  
Jean-Noël Vittaut, Patrick Gallinari ........................................ 336

**Relevance Feedback Track**

Relevance Feedback for Structural Query Expansion  
Ralf Schenkel, Martin Theobald ............................................. 344

**Natural Language Query Track**

NLPX at INEX 2005  
Alan Woodley, Shlomo Geva ............................................. 358

From Natural Language to NEXI, an Interface for INEX 2005 Queries  
Xavier Tannier .......................................................................... 373

**Heterogeneous Track**

Processing Heterogeneous Collections in XML Information Retrieval  
Maria Izabel Menezes Azevedo, Klérisson Vinícius Ribeiro Paixão,  
Diego Vinícius Castro Pereira .................................................. 388

**Interactive Track**

The Interactive Track at INEX 2005  
Birger Larsen, Saadia Malik, Anastasios Tombros ....................... 398

What Do Users Think of an XML Element Retrieval System?  
Jaap Kamps, Börkur Sigurbjörnsson ........................................... 411

Users Interaction with the Hierarchically Structured Presentation in XML Document Retrieval  
Heesop Kim, Heejung Son ....................................................... 422

**Document Mining Track**

XML Documents Clustering by Structures  
Richi Nayak, Sumei Xu .......................................................... 432
<table>
<thead>
<tr>
<th>Title</th>
<th>Authors</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>A Flexible Structured-Based Representation for XML Document Mining</td>
<td>Anne-Marie Vercoustre, Mounir Fegas, Saba Gul, Yves Lechevallier</td>
<td>443</td>
</tr>
<tr>
<td>Sequential Pattern Mining for Structure-Based XML Document Classification</td>
<td>Calin Garboni, Florent Masseglia, Brigitte Trousse</td>
<td>458</td>
</tr>
<tr>
<td>Transforming XML Trees for Efficient Classification and Clustering</td>
<td>Laurent Candillier, Isabelle Tellier, Fabien Torre</td>
<td>469</td>
</tr>
<tr>
<td>Multimedia Track</td>
<td></td>
<td></td>
</tr>
<tr>
<td>INEX 2005 Multimedia Track</td>
<td>Roelof van Zwol, Gabriella Kazai, Mounia Lalmas</td>
<td>497</td>
</tr>
<tr>
<td>Integrating Text Retrieval and Image Retrieval in XML Document Searching</td>
<td>D. Tjondronegoro, J. Zhang, J. Gu, A. Nguyen, S. Geva</td>
<td>511</td>
</tr>
<tr>
<td>Combining Image and Structured Text Retrieval</td>
<td>D.N.F. Awang Iskandar, Jovan Pehcevski, James A. Thom, S.M.M. Tahaghoghi</td>
<td>525</td>
</tr>
<tr>
<td>Multimedia Strategies for $B^3$-SDR, Based on Principal Component Analysis</td>
<td>Roelof van Zwol</td>
<td>540</td>
</tr>
<tr>
<td>Author Index</td>
<td></td>
<td>555</td>
</tr>
</tbody>
</table>