

*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  
Andrew Trotman (Eds.)

# Comparative Evaluation of XML Information Retrieval Systems

5th International Workshop of the Initiative  
for the Evaluation of XML Retrieval, INEX 2006  
Dagstuhl Castle, Germany, December 17-20, 2006  
Revised and Selected Papers

## Volume Editors

Norbert Fuhr

Department of Informatics

University of Duisburg-Essen, 47048 Duisburg, Germany

E-mail: [norbert.fuhr@uni-due.de](mailto:norbert.fuhr@uni-due.de)

Mounia Lalmas

Department of Computer Science

Queen Mary, University of London, London, UK

E-mail: [mounia@dcs.qmul.ac.uk](mailto:mounia@dcs.qmul.ac.uk)

Andrew Trotman

Department of Computer Science

University of Otago

Dunedin, New Zealand

E-mail: [andrew@cs.otago.ac.nz](mailto:andrew@cs.otago.ac.nz)

Library of Congress Control Number: 2007932681

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 3-540-73887-8 Springer Berlin Heidelberg New York

ISBN-13 978-3-540-73887-9 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

[springer.com](http://springer.com)

© Springer-Verlag Berlin Heidelberg 2007

Printed in Germany

Typesetting: Camera-ready by author, data conversion by Scientific Publishing Services, Chennai, India

Printed on acid-free paper SPIN: 12098521 06/3180 5 4 3 2 1 0

# Preface

Welcome to the fifth workshop of the Initiative for the Evaluation of XML Retrieval (INEX)!

Now in its fifth year, INEX is an established evaluation forum for XML information retrieval (IR), with over 80 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 plays an increasingly important role in many information access systems (e.g., digital libraries, Web, intranet) where content is 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.

In total, nine research tracks were included in INEX 2006, which studied different aspects of XML information access: Ad-hoc, Interactive, Use Case, Multimedia, Relevance Feedback, Heterogeneous, Document Mining, Natural Language Processing, and Entity Ranking. The Use Case and Entity Ranking tracks were new in 2006. The consolidation of the existing tracks, and the expansion to new areas offered by the two new tracks, allowed INEX to grow in reach.

The aim of the INEX 2006 workshop was to bring together researchers in the field of XML IR who participated in the INEX 2006 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 concluded the results of this large-scale effort, summarized and addressed the encountered issues and devised a work plan for the future evaluation of XML retrieval systems.

INEX was funded by the DELOS Network of Excellence on Digital Libraries, to which we are very thankful. 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 individuals for their contributions.

March 2007

Norbert Fuhr  
Mounia Lalmas  
Andrew Trotman

# Organization

## Project Leaders

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

## Contact Persons

Saadia Malik	University of Duisburg-Essen, Germany
Zoltn Szlavik	Queen Mary, University of London, UK

## Wikipedia Document Collection

Ludovic Denoyer	Universite Paris 6, France
Martin Theobald	Max Planck Institute for Informatics, Germany

## Use Case Studies

Andrew Trotman	University of Otago, New Zealand
Nils Pharo	Oslo University College, Norway

## Topic Format Specification

Andrew Trotman	University of Otago, New Zealand
Birger Larsen	Royal School of LIS, Denmark

## 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
---------------------	--------------------------------------

## Metrics

Gabriella Kazai	Microsoft Research Cambridge, UK
Stephen Robertson	Microsoft Research Cambridge, UK
Paul Ogilvie	Carnegie Mellon University , USA

## Relevance Feedback Task

Yosi Mass	IBM Research Lab, Israel
Ralf Schenkel	Max Planck Institute for Informatics, Germany

## Natural Query Language Task

Shlomo Geva	Queensland University of Technology, Australia
Xavier Tannier	XEROX, France

## Heterogeneous Collection Track

Ingo Frommholz	University of Duisburg-Essen, Germany
Ray Larson	University of California, Berkeley, USA

## Interactive Track

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

## Document Mining Track

Ludovic Denoyer	Universite Paris 6, France
Anne-Marie Vercoustre	Inria-Rocquencourt, France
Patrick Gallinari	Universite Paris 6, France

## XML Multimedia Track

Roelof van Zwol	Yahoo! Research, Spain
Thijs Westerveld	CWI, The Netherlands

## XML Entity Search Track

Arjen de Vries	CWI, The Netherlands
Nick Craswell	Microsoft Research Cambridge, UK

# Table of Contents

## Methodology

Overview of INEX 2006 .....	1
<i>Saadia Malik, Andrew Trotman, Mounia Lalmas, and Norbert Fuhr</i>	
The Wikipedia XML Corpus .....	12
<i>Ludovic Denoyer and Patrick Gallinari</i>	
INEX 2006 Evaluation Measures .....	20
<i>Mounia Lalmas, Gabriella Kazai, Jaap Kamps, Jovan Pehcevski, Benjamin Piwowarski, and Stephen Robertson</i>	
Choosing an Ideal Recall-Base for the Evaluation of the Focused Task: Sensitivity Analysis of the XCG Evaluation Measures .....	35
<i>Gabriella Kazai</i>	

## Ad Hoc Track

A Method of Preferential Unification of Plural Retrieved Elements for XML Retrieval Task .....	45
<i>Hiroki Tanioka</i>	
CISR at INEX 2006 .....	57
<i>Wei Lu, Stephen Robertson, and Andrew Macfarlane</i>	
Compact Representations in XML Retrieval .....	64
<i>Fang Huang, Stuart Watt, David Harper, and Malcolm Clark</i>	
CSIRO's Participation in INEX 2006 .....	73
<i>Alexander Krumpholz and David Hawking</i>	
Dynamic Element Retrieval in a Semi-structured Collection .....	82
<i>Carolyn J. Crouch, Donald B. Crouch, Murthy Ganapathibhotla, and Vishal Bakshi</i>	
Efficient, Effective and Flexible XML Retrieval Using Summaries .....	89
<i>M.S. Ali, Mariano Consens, Xin Gu, Yaron Kanza, Flavio Rizzolo, and Raquel Stasiu</i>	
Evaluating Structured Information Retrieval and Multimedia Retrieval Using PF/Tijah .....	104
<i>Thijs Westerveld, Henning Rode, Roel van Os, Djoerd Hiemstra, Georgina Ramírez, Vojkan Mihajlović, and Arjen P. de Vries</i>	

EXTIRP: Baseline Retrieval from Wikipedia .....	115
<i>Miro Lehtonen and Antoine Doucet</i>	
Filtering and Clustering XML Retrieval Results .....	121
<i>Jaap Kamps, Marijn Koolen, and Börkur Sigurbjörnsson</i>	
GPX - Gardens Point XML IR at INEX 2006 .....	137
<i>Shlomo Geva</i>	
IBM HRL at INEX 06 .....	151
<i>Yosi Mass</i>	
Indexing “Reading Paths” for a Structured Information Retrieval at INEX 2006 .....	160
<i>Mathias Géry</i>	
Influence Diagrams and Structured Retrieval: Garnata Implementing the SID and CID Models at INEX’06 .....	165
<i>Luis M. de Campos, Juan M. Fernández-Luna, Juan F. Huete, and Alfonso E. Romero</i>	
Information Theoretic Retrieval with Structured Queries and Documents .....	178
<i>Claudio Carpineto, Giovanni Romano, and Caterina Caracciolo</i>	
SIRIUS XML IR System at INEX 2006: Approximate Matching of Structure and Textual Content .....	185
<i>Eugen Popovici, Gildas Ménier, and Pierre-François Marteau</i>	
Structured Content-Only Information Retrieval Using Term Proximity and Propagation of Title Terms .....	200
<i>Michel Beigbeder</i>	
Supervised and Semi-supervised Machine Learning Ranking .....	213
<i>Jean-Noël Vittaut and Patrick Gallinari</i>	
The University of Kaiserslautern at INEX 2006 .....	223
<i>Philipp Dopichaj</i>	
TopX – AdHoc Track and Feedback Task .....	233
<i>Martin Theobald, Andreas Broschart, Ralf Schenkel, Silvana Solomon, and Gerhard Weikum</i>	
Tuning and Evolving Retrieval Engine by Training on Previous INEX Testbeds .....	243
<i>Gilles Hubert</i>	
Using Language Models and the HITS Algorithm for XML Retrieval ...	253
<i>Benny Kimelfeld, Eitan Kovacs, Yehoshua Sagiv, and Dan Yahav</i>	



Using Topic Shifts in XML Retrieval at INEX 2006 .....	261
<i>Elham Ashoori and Mounia Lalmas</i>	

XSee: Structure Xposed.....	271
<i>Roelof van Zwol and Wouter Weerkamp</i>	

## Natural Language Processing Track

Shallow Parsing of INEX Queries .....	284
<i>Haïfa Zargayouna, Victor Rosas, and Sylvie Salotti</i>	

Using Rich Document Representation in XML Information Retrieval ...	294
<i>Fahimeh Raja, Mostafa Keikha, Masued Rahgozar, and Farhad Oroumchian</i>	

NLPX at INEX 2006 .....	302
<i>Alan Woodley and Shlomo Geva</i>	

## Heterogeneous Collection Track

The Heterogeneous Collection Track at INEX 2006 .....	312
<i>Ingo Frommholz and Ray Larson</i>	

Probabilistic Retrieval Approaches for Thorough and Heterogeneous XML Retrieval .....	318
<i>Ray R. Larson</i>	

## Multimedia Track

The INEX 2006 Multimedia Track .....	331
<i>Thijs Westerveld and Roelof van Zwol</i>	

Fusing Visual and Textual Retrieval Techniques to Effectively Search Large Collections of Wikipedia Images .....	345
<i>C. Lau, D. Tjondronegoro, J. Zhang, S. Geva, and Y. Liu</i>	

Social Media Retrieval Using Image Features and Structured Text .....	358
<i>D.N.F. Awang Iskandar, Jovan Pehcevski, James A. Thom, and S.M.M. Tahaghoghi</i>	

XFIRM at INEX 2006. Ad-Hoc, Relevance Feedback and MultiMedia Tracks .....	373
<i>Lobna Hlaoua, Mouna Torjmen, Karen Pinel-Sauvagnat, and Mohand Boughanem</i>	

## Interactive Track

The Interactive Track at INEX 2006 .....	387
<i>Saadia Malik, Anastasios Tombros, and Birger Larsen</i>	

## Use Case Track

XML-IR Users and Use Cases .....	400
<i>Andrew Trotman, Nils Pharo, and Miro Lehtonen</i>	
A Taxonomy for XML Retrieval Use Cases .....	413
<i>Miro Lehtonen, Nils Pharo, and Andrew Trotman</i>	
What XML-IR Users May Want .....	423
<i>Alan Woodley, Shlomo Geva, and Sylvia L. Edwards</i>	

## Document Track

Report on the XML Mining Track at INEX 2005 and INEX 2006 .....	432
<i>Ludovic Denoyer, Patrick Gallinari, and Anne-Marie Vercoustre</i>	
Classifying XML Documents Based on Structure/Content Similarity....	444
<i>Guangming Xing, Jinhua Guo, and Zhonghang Xia</i>	
Document Mining Using Graph Neural Network .....	458
<i>S.L. Yong, M. Hagenbuchner, A.C. Tsoi, F. Scarselli, and M. Gori</i>	
Evaluating the Performance of XML Document Clustering by Structure Only .....	473
<i>Tien Tran and Richi Nayak</i>	
FAT-CAT: Frequent Attributes Tree Based Classification.....	485
<i>Jeroen De Knijf</i>	
Unsupervised Classification of Text-Centric XML Document Collections .....	497
<i>Antoine Doucet and Miro Lehtonen</i>	
XML Document Mining Using Contextual Self-organizing Maps for Structures .....	510
<i>M. Kc, M. Hagenbuchner, A.C. Tsoi, F. Scarselli, A. Sperduti, and M. Gori</i>	
XML Document Transformation with Conditional Random Fields .....	525
<i>Rémi Gilleron, Florent Jousse, Isabelle Tellier, and Marc Tommasi</i>	
XML Structure Mapping.....	540
<i>Francis Maes, Ludovic Denoyer, and Patrick Gallinari</i>	
<b>Author Index .....</b>	<b>553</b>