

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

Alfred Kobsa

University of California, Irvine, CA, 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

Gerhard Weikum

Max-Planck Institute of Computer Science, Saarbruecken, Germany

Norbert Fuhr Jaap Kamps Mounia Lalmas
Andrew Trotman (Eds.)

Focused Access to XML Documents

6th International Workshop of the Initiative
for the Evaluation of XML Retrieval, INEX 2007
Dagstuhl Castle, Germany, December 17-19, 2007
Revised and Selected Papers

Volume Editors

Norbert Fuhr

University of Duisburg-Essen, Department of Computational
and Cognitive Sciences, 47048 Duisburg, Germany

E-mail: norbert.fuhr@uni-due.de

Jaap Kamps

University of Amsterdam, Archives and Information Studies/Humanities
1012 Amsterdam, The Netherlands

E-mail: kamps@science.uva.nl

Mounia Lalmas

University of London, Department of Computer Science
Queen Mary, E1 4NS, London, United Kingdom

E-mail: mounia@dcs.qmul.ac.uk

Andrew Trotman

University of Otago, Department of Computer Science
9015 Dunedin, New Zealand

E-mail: andrew@cs.otago.ac.nz

Library of Congress Control Number: 2008934300

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 3-540-85901-2 Springer Berlin Heidelberg New York

ISBN-13 978-3-540-85901-7 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

springer.com

© Springer-Verlag Berlin Heidelberg 2008

Printed in Germany

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

Printed on acid-free paper SPIN: 12522002 06/3180 5 4 3 2 1 0

Preface

Welcome to the proceedings of the 6th workshop of the Initiative for the Evaluation of XML Retrieval (INEX)! Now in its sixth year, INEX has become an established evaluation forum for XML information retrieval (IR), with over 100 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 is playing an increasingly important role in many information access systems (e.g., digital libraries, web, intranet) where content is becoming 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.

The aim of the INEX 2007 workshop was to bring together researchers in the field of XML IR who participated in the INEX 2007 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 brought together the results of this large-scale effort, summarized and addressed the issues encountered, and devised a work plan for the future evaluation of XML retrieval systems.

In total seven research tracks were included in INEX 2007. These studied different aspects of XML information access: ad-hoc, document mining, multimedia, heterogeneous, entity ranking, book search, and link-the-wiki. The consolidation of the existing tracks, and the expansion to new areas offered by the new tracks has enabled INEX to extend its scope. This volume contains 37 papers selected from 50 submitted ones (74% acceptance rate). Each paper was peer-reviewed.

INEX is funded by the DELOS Network of Excellence on Digital Libraries, to which we are very thankful. We thank Schloss Dagstuhl (Leibniz Center for Informatics) for housing the INEX workshop in its unique and inspiring atmosphere, and Springer for publishing the results of INEX 2007 in these final proceedings. 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 people for their contribution.

May 2008

Norbert Fuhr
Jaap Kamps
Mounia Lalmas
Andrew Trotman

Organization

Project Leaders

Norbert Fuhr	University of Duisburg-Essen, Germany
Mounia Lalmas	Queen Mary, University of London, UK
Andrew Trotman	University of Otago, New Zealand

Contact People

Saadia Malik	University of Duisburg-Essen, Germany
Zoltán Szilávik	Queen Mary, University of London, UK

Wikipedia Document Collection

Ludovic Denoyer	University Paris 6, France
-----------------	----------------------------

Document Exploration

Ralf Schenkel	Max-Planck-Institut für Informatik, Germany
Martin Theobald	Stanford University, USA

Topic Format Specification

Birger Larsen	Royal School of Library and Information Science, Denmark
Andrew Trotman	University of Otago, New Zealand

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
---------------------	--------------------------------------

Effectiveness Measures

Gabriella Kazai	Microsoft Research Cambridge, UK
Benjamin Piwowarski	Yahoo! Research Latin America, Chile

Jaap Kamps	University of Amsterdam, The Netherlands
Jovan Pehcevski	INRIA Rocquencourt, France
Stephen Robertson	Microsoft Research Cambridge, UK

Document Mining Track

Ludovic Denoyer	University Paris 6, France
Patrick Gallinari	University Paris 6, France

Multimedia Track

Thijs Westerveld	CWI, The Netherlands
Theodora Tsikrika	CWI, The Netherlands

Entity Ranking Track

Arjen de Vries	CWI, The Netherlands
Nick Craswell	Microsoft Research Cambridge, UK
Mounia Lalmas	Queen Mary, University of London, UK
James A. Thom	RMIT University, Australia
Anne-Marie Vercoustre	INRIA Rocquencourt, France

Link the Wiki Track

Shlomo Geva	Queensland University of Technology, Australia
Andrew Trotman	University of Otago, New Zealand

Book Search

Gabriella Kazai	Microsoft Research Cambridge, UK
Antoine Doucet	University of Caen, France

Table of Contents

Ad Hoc Track

Overview of the INEX 2007 Ad Hoc Track	1
<i>Norbert Fuhr, Jaap Kamps, Mounia Lalmas, Saadia Malik, and Andrew Trotman</i>	
INEX 2007 Evaluation Measures	24
<i>Jaap Kamps, Jovan Pehcevski, Gabriella Kazai, Mounia Lalmas, and Stephen Robertson</i>	
XML Retrieval by Improving Structural Relevance Measures Obtained from Summary Models	34
<i>M.S. Ali, Mariano P. Consens, and Shahan Khatchadourian</i>	
TopX @ INEX 2007	49
<i>Andreas Broschart, Ralf Schenkel, Martin Theobald, and Gerhard Weikum</i>	
The Garnata Information Retrieval System at INEX'07	57
<i>Luis M. de Campos, Juan M. Fernández-Luna, Juan F. Huete, Carlos Martín-Dancausa, and Alfonso E. Romero</i>	
Dynamic Element Retrieval in the Wikipedia Collection	70
<i>Carolyn J. Crouch, Donald B. Crouch, Nachiket Kamat, Vikram Malik, and Aditya Mone</i>	
The Simplest XML Retrieval Baseline That Could Possibly Work	80
<i>Philipp Dopichaj</i>	
Using Language Models and Topic Models for XML Retrieval	94
<i>Fang Huang</i>	
UJM at INEX 2007: Document Model Integrating XML Tags	103
<i>Mathias Géry, Christine Largeron, and Franck Thollard</i>	
Phrase Detection in the Wikipedia	115
<i>Miro Lehtonen and Antoine Doucet</i>	
Indian Statistical Institute at INEX 2007 Adhoc Track: VSM Approach	122
<i>Sukomal Pal and Mandar Mitra</i>	
A Fast Retrieval Algorithm for Large-Scale XML Data	129
<i>Hiroki Tanioka</i>	

LIG at INEX 2007 Ad Hoc Track: Using Collectionlinks as Context	138
<i>Delphine Verbyst and Philippe Mulhem</i>	

Book Search Track

Overview of the INEX 2007 Book Search Track (BookSearch'07)	148
<i>Gabriella Kazai and Antoine Doucet</i>	
Logistic Regression and EVIs for XML Books and the Heterogeneous Track	162
<i>Ray R. Larson</i>	
CMIC at INEX 2007: Book Search Track	175
<i>Walid Magdy and Kareem Darwish</i>	

XML-Mining Track

Clustering XML Documents Using Closed Frequent Subtrees: A Structural Similarity Approach	183
<i>Sangeetha Kutty, Tien Tran, Richi Nayak, and Yuefeng Li</i>	
Probabilistic Methods for Structured Document Classification at INEX'07	195
<i>Luis M. de Campos, Juan M. Fernández-Luna, Juan F. Huete, and Alfonso E. Romero</i>	
Efficient Clustering of Structured Documents Using Graph Self-Organizing Maps	207
<i>Markus Hagenbuchner, Ah Chung Tsoi, Alessandro Sperduti, and Milly Kc</i>	
Document Clustering Using Incremental and Pairwise Approaches	222
<i>Tien Tran, Richi Nayak, and Peter Bruza</i>	
XML Document Classification Using Extended VSM	234
<i>Jianwu Yang and Fudong Zhang</i>	

Entity Ranking Track

Overview of the INEX 2007 Entity Ranking Track	245
<i>Arjen P. de Vries, Anne-Marie Vercoustre, James A. Thom, Nick Craswell, and Mounia Lalmas</i>	
L3S at INEX 2007: Query Expansion for Entity Ranking Using a Highly Accurate Ontology	252
<i>Gianluca Demartini, Claudiu S. Firan, and Tereza Iofciu</i>	

Entity Ranking Based on Category Expansion	264
<i>Janne Jämsen, Turkka Näppilä, and Paavo Arvola</i>	
Entity Ranking from Annotated Text Collections Using Multitype Topic Models	279
<i>Hitohiro Shiozaki and Koji Eguchi</i>	
An n-Gram and Initial Description Based Approach for Entity Ranking Track	293
<i>Meenakshi Sundaram Murugesan and Saswati Mukherjee</i>	
Structured Document Retrieval, Multimedia Retrieval, and Entity Ranking Using PF/Tijah	306
<i>Theodora Tsikrika, Pavel Serdyukov, Henning Rode, Thijs Westerveld, Robin Aly, Djoerd Hiemstra, and Arjen P. de Vries</i>	
Using Wikipedia Categories and Links in Entity Ranking	321
<i>Anne-Marie Vercoustre, Jovan Pehcevski, and James A. Thom</i>	
Integrating Document Features for Entity Ranking	336
<i>Jianhan Zhu, Dawei Song, and Stefan Rüger</i>	

Interactive Track

A Comparison of Interactive and Ad-Hoc Relevance Assessments	348
<i>Birger Larsen, Saadia Malik, and Anastasios Tombros</i>	
Task Effects on Interactive Search: The Query Factor	359
<i>Elaine G. Toms, Heather O'Brien, Tayze Mackenzie, Chris Jordan, Luanne Freund, Sandra Toze, Emilie Dawe, and Alexandra MacNutt</i>	

Link-the-Wiki Track

Overview of INEX 2007 Link the Wiki Track	373
<i>Darren Wei Che Huang, Yue Xu, Andrew Trotman, and Shlomo Geva</i>	
Using and Detecting Links in Wikipedia	388
<i>Khairun Nisa Fachry, Jaap Kamps, Marijn Koolen, and Junte Zhang</i>	
GPX: Ad-Hoc Queries and Automated Link Discovery in the Wikipedia	404
<i>Shlomo Geva</i>	
University of Waterloo at INEX2007: Adhoc and Link-the-Wiki Tracks	417
<i>Kelly Y. Itakura and Charles L.A. Clarke</i>	

Wikipedia *Ad Hoc* Passage Retrieval and Wikipedia Document
Linking 426
 Dylan Jenkinson and Andrew Trotman

Multimedia Track

The INEX 2007 Multimedia Track 440
 Theodora Tsikrika and Thijs Westerveld

Author Index 455