

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

Microsoft Research, Cambridge, 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

Shlomo Geva Jaap Kamps
Andrew Trotman (Eds.)

Advances in Focused Retrieval

7th International Workshop of the Initiative
for the Evaluation of XML Retrieval, INEX 2008
Dagstuhl Castle, Germany, December 15-18, 2008
Revised and Selected Papers



Springer

Volume Editors

Shlomo Geva
Queensland University of Technology
Faculty of Science and Technology
GPO Box 2434, Brisbane Qld 4001, Australia
E-mail: s.geva@qut.edu.au

Jaap Kamps
University of Amsterdam
Archives and Information Studies/Humanities
Turfdraagsterpad 9, 1012 XT Amsterdam, The Netherlands
E-mail: kamps@uva.nl

Andrew Trotman
University of Otago
Department of Computer Science
P.O. Box 56, Dunedin 9054, New Zealand
E-mail: andrew@cs.otago.ac.nz

Library of Congress Control Number: 2009933190

CR Subject Classification (1998): H.3, H.3.3, H.3.4, H.2.8, H.2.3, H.2.4, E.1

LNCS Sublibrary: SL 3 – Information Systems and Application, incl. Internet/Web and HCI

ISSN 0302-9743
ISBN-10 3-642-03760-7 Springer Berlin Heidelberg New York
ISBN-13 978-3-642-03760-3 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.com

© Springer-Verlag Berlin Heidelberg 2009
Printed in Germany

Typesetting: Camera-ready by author, data conversion by Scientific Publishing Services, Chennai, India
Printed on acid-free paper SPIN: 12727117 06/3180 5 4 3 2 1 0

Foreword

I write with pleasure this foreword to the proceedings of the 7th workshop of the Initiative for the Evaluation of XML Retrieval (INEX). The increased adoption of XML as the standard for representing a document structure has led to the development of retrieval systems that are aimed at effectively accessing XML documents. Providing effective access to large collections of XML documents is therefore a key issue for the success of these systems. INEX aims to provide the necessary methodological means and worldwide infrastructures for evaluating how good XML retrieval systems are.

Since its launch in 2002, INEX has grown both in terms of number of participants and its coverage of the investigated retrieval tasks and scenarios. In 2002, INEX started with 49 registered participating organizations, whereas this number was more than 100 for 2008. In 2002, there was one main track, concerned with the ad hoc retrieval task, whereas in 2008, seven tracks in addition to the main ad hoc track were investigated, looking at various aspects of XML retrieval, from book search to entity ranking, including interaction aspects.

INEX follows the predominant approach in information retrieval of evaluating retrieval approaches using a test collection constructed specifically for that purpose and associated effectiveness measures. A test collection usually consists of a set of documents, user requests (topics), and relevance assessments, which specify the set of “right answers” for the requests. Throughout the years, INEX faced a range of challenges regarding the evaluation of XML retrieval systems, as the consideration of the structure led to many complex issues, which were not always identified at the beginning (e.g., the issue of overlap and a counting-based effectiveness measure, the difficulty in consistently assessing elements using a four-graded two-dimensional scale). In addition, limited funding was available, for example, to pay assessors. This led to a research problem in itself, namely, how to elicit quality assessments in order for the test collections to be reusable. As a result, different theories and methods for the evaluation of XML retrieval were developed and tested at INEX, e.g., the definition of relevance, a plethora of effectiveness measures, leading to a now-stable evaluation setup and a rich history of learned lessons. This is now allowing researchers worldwide to make further progress in XML retrieval, including the investigation of other research questions in XML retrieval, for instance, efficiency in 2008.

What I have greatly enjoyed with INEX is the people working together not only to develop approaches for XML retrieval, but also methodologies for evaluating XML retrieval. Many of the people that actually joined INEX to test their XML retrieval approaches got *hooked* on the problem of properly evaluating XML retrieval approaches. Three of them are the current INEX organizers, Shlomo Geva, Jaap Kamps, and Andrew Trotman. I am very glad that they got hooked, as they are dedicated and, even more importantly, enthusiastic people

with extensive expertise in both building XML retrieval systems and evaluating them. INEX is in very safe hands with them. Looking at the current proceedings, I am delighted to see so many excellent works from people from all over the world. Many of them met at the annual workshop, and I heard it was a great success. Well done to all.

April 2009

Mounia Lalmas

Preface

Welcome to the 7th workshop of the Initiative for the Evaluation of XML Retrieval (INEX)! Now, in its seventh year, INEX is one of the established evaluation forums in information retrieval (IR), with 150 organizations worldwide registering and over 50 groups participating actively in the different tracks. INEX aims to provide an infrastructure, in the form of a large structured test collection and appropriate scoring methods, for the evaluation of focused retrieval.

Information on the Web is a mixture of text, multimedia, and metadata, with a clear internal structure, usually formatted according to the eXtensible Markup Language (XML) standard, or another related W3C standard. 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.

INEX 2008 was an exciting year for INEX, and brought a lot of changes. Seven research tracks were included, which studied different aspects of focused information access: Ad Hoc, Book, Efficiency, Entity Ranking, Interactive (iTrack), Link the Wiki, and XML Mining. The aim of the INEX 2008 workshop was to bring together researchers who participated in the INEX 2008 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 issues encountered, and devised a work plan for the future evaluation of XML retrieval systems. These proceedings report the final results of INEX 2008. We accepted a total of 49 out of 53 papers, yielding a 92% acceptance rate.

This was also the seventh INEX Workshop to be held at the *Schloss Dagstuhl – Leibniz Center for Informatics*, providing a unique setting where informal interaction and discussion occurs naturally and frequently. This has been essential to the growth of INEX over the years, and we feel honored and privileged that *Dagstuhl* housed the INEX 2008 Workshop. Finally, INEX was run for, but especially by, the participants. It was a result of tracks and tasks suggested by participants, topics created by participants, systems built by participants, and relevance judgments provided by participants. So the main thank you goes to each of these individuals!

April 2009

Shlomo Geva
Jaap Kamps
Andrew Trotman

Organization

Steering Committee

Charlie Clarke	University of Waterloo, Canada
Norbert Fuhr	University of Duisburg-Essen, Germany
Shlomo Geva	Queensland University of Technology, Australia
Jaap Kamps	University of Amsterdam, The Netherlands
Mounia Lalmas	University of Glasgow, UK
Stephen Robertson	Microsoft Research Cambridge, UK
Andrew Trotman	University of Otago, New Zealand
Ellen Voorhees	NIST, USA

Chairs

Shlomo Geva	Queensland University of Technology, Australia
Jaap Kamps	University of Amsterdam, The Netherlands
Andrew Trotman	University of Otago, New Zealand

Track Organizers

Ad Hoc

Shlomo Geva	General, Queensland University of Technology, Australia
Jaap Kamps	General, University of Amsterdam, The Netherlands
Andrew Trotman	General, University of Otago, New Zealand
Ludovic Denoyer	Document Collection, University Paris 6, France
Ralf Schenkel	Document Exploration, Max-Planck-Institut für Informatik, Germany
Martin Theobald	Document Exploration, Stanford University, USA

Book

Antoine Doucet	University of Caen, France
Gabriella Kazai	Microsoft Research Limited, Cambridge, UK
Monica Landoni	University of Strathclyde, UK

Efficiency

Ralf Schenkel
Martin Theobald

Max-Planck-Institut für Informatik, Germany
Stanford University, USA

Entity Ranking

Gianluca Demartini
Tereza Iofciu
Arjen de Vries
Jianhan Zhu

L3S, Leibniz Universität Hannover, Germany
L3S, Leibniz Universität Hannover, Germany
CWI, The Netherlands
University College London, UK

Interactive (iTrack)

Nisa Fachry
Ragnar Nordlie
Nils Pharo

University of Amsterdam, The Netherlands
Oslo University College, Norway
Oslo University College, Norway

Link the Wiki

Shlomo Geva

Wei-Che (Darren) Huang

Andrew Trotman

Queensland University of Technology,
Australia
Queensland University of Technology,
Australia
University of Otago, New Zealand

XML Mining

Ludovic Denoyer
Patrick Gallinari

University Paris 6, France
University Paris 6, France

Table of Contents

Ad Hoc Track

Overview of the INEX 2008 Ad Hoc Track	1
<i>Jaap Kamps, Shlomo Geva, Andrew Trotman, Alan Woodley, and Marijn Koolen</i>	
Experiments with Proximity-Aware Scoring for XML Retrieval at INEX 2008	29
<i>Andreas Broschart, Ralf Schenkel, and Martin Theobald</i>	
Finding Good Elements for Focused Retrieval	33
<i>Carolyn J. Crouch, Donald B. Crouch, Salil Bapat, Sarika Mehta, and Darshan Paranjape</i>	
New Utility Models for the Garnata Information Retrieval System at INEX'08	39
<i>Luis M. de Campos, Juan M. Fernández-Luna, Juan F. Huete, Carlos Martín-Dancausa, and Alfonso E. Romero</i>	
UJM at INEX 2008: Pre-impacting of Tags Weights	46
<i>Mathias Géry, Christine Largeron, and Franck Thollard</i>	
Use of Multiword Terms and Query Expansion for Interactive Information Retrieval	54
<i>Fidelia Ibekwe-SanJuan and Eric SanJuan</i>	
Enhancing Keyword Search with a Keyphrase Index	65
<i>Miro Lehtonen and Antoine Doucet</i>	
CADIAL Search Engine at INEX	71
<i>Jure Mijić, Marie-Francine Moens, and Bojana Dalbelo Bašić</i>	
Indian Statistical Institute at INEX 2008 Adhoc Track	79
<i>Sukomal Pal, Mandar Mitra, Debasis Ganguly, Samarendra Maiti, Ayan Bandyopadhyay, Aparajita Sen, and Sukanya Mitra</i>	
Using Collectionlinks and Documents as Context for INEX 2008	87
<i>Delphine Verbyst and Philippe Mulhem</i>	
SPIRIX: A Peer-to-Peer Search Engine for XML-Retrieval	97
<i>Judith Winter and Oswald Drobnik</i>	

Book Track

Overview of the INEX 2008 Book Track	106
<i>Gabriella Kazai, Antoine Doucet, and Monica Landoni</i>	
XRCE Participation to the Book Structure Task	124
<i>Hervé Déjean and Jean-Luc Meunier</i>	
University of Waterloo at INEX 2008: Adhoc, Book, and Link-the-Wiki Tracks	132
<i>Kelly Y. Itakura and Charles L.A. Clarke</i>	
The Impact of Document Level Ranking on Focused Retrieval	140
<i>Jaap Kamps and Marijn Koolen</i>	
Adhoc and Book XML Retrieval with Cheshire	152
<i>Ray R. Larson</i>	
Book Layout Analysis: TOC Structure Extraction Engine	164
<i>Bodin Dresevic, Aleksandar Uzelac, Bogdan Radakovic, and Nikola Todic</i>	
The Impact of Query Length and Document Length on Book Search Effectiveness	172
<i>Mingfang Wu, Falk Scholer, and James A. Thom</i>	

Efficiency Track

Overview of the INEX 2008 Efficiency Track	179
<i>Martin Theobald and Ralf Schenkel</i>	
Exploiting User Navigation to Improve Focused Retrieval	192
<i>M.S. Ali, Mariano P. Consens, Bassam Helou, and Shahan Khatchadourian</i>	
Efficient XML and Entity Retrieval with PF/Tijah: CWI and University of Twente at INEX'08	207
<i>Henning Rode, Djoerd Hiemstra, Arjen de Vries, and Pavel Serdyukov</i>	
Pseudo Relevance Feedback Using Fast XML Retrieval	218
<i>Hiroki Tanioka</i>	
TopX 2.0 at the INEX 2008 Efficiency Track: A (Very) Fast Object-Store for Top-k-Style XML Full-Text Search	224
<i>Martin Theobald, Mohammed AbuJarour, and Ralf Schenkel</i>	
Aiming for Efficiency by Detecting Structural Similarity	237
<i>Judith Winter, Nikolay Jeliazkov, and Gerold Kühne</i>	

Entity Ranking Track

Overview of the INEX 2008 Entity Ranking Track	243
<i>Gianluca Demartini, Arjen de Vries, Tereza Iofciu, and Jianhan Zhu</i>	
L3S at INEX 2008: Retrieving Entities Using Structured Information ...	253
<i>Nick Craswell, Gianluca Demartini, Julien Gaugaz, and Tereza Iofciu</i>	
Adapting Language Modeling Methods for Expert Search to Rank Wikipedia Entities	264
<i>Jiepu Jiang, Wei Lu, Xianqian Rong, and Yangyan Gao</i>	
Finding Entities in Wikipedia Using Links and Categories	273
<i>Rianne Kaptein and Jaap Kamps</i>	
Topic Difficulty Prediction in Entity Ranking	280
<i>Anne-Marie Vercoustre, Jovan Pehcevski, and Vladimir Naumovski</i>	
A Generative Language Modeling Approach for Ranking Entities	292
<i>Wouter Weerkamp, Krisztian Balog, and Edgar Meij</i>	

Interactive Track

Overview of the INEX 2008 Interactive Track	300
<i>Nils Pharo, Ragnar Nordlie, and Khairun Nisa Fachry</i>	

Link the Wiki Track

Overview of the INEX 2008 Link the Wiki Track	314
<i>Wei Che (Darren) Huang, Shlomo Geva, and Andrew Trotman</i>	
Link-the-Wiki: Performance Evaluation Based on Frequent Phrases	326
<i>Mao-Lung (Edward) Chen, Richi Nayak, and Shlomo Geva</i>	
CMIC@INEX 2008: Link-the-Wiki Track	337
<i>Kareem Darwish</i>	
Stealing Anchors to Link the Wiki	343
<i>Philipp Dopichaj, Andre Skusa, and Andreas Heß</i>	
Context Based Wikipedia Linking	354
<i>Michael Granitzer, Christin Seifert, and Mario Zechner</i>	
Link Detection with Wikipedia	366
<i>Jiyin He</i>	
Wikisearching and Wikilinking	374
<i>Dylan Jenkinson, Kai-Cheung Leung, and Andrew Trotman</i>	

CSIR at INEX 2008 Link-the-Wiki Track	389
<i>Wei Lu, Dan Liu, and Zhenzhen Fu</i>	
A Content-Based Link Detection Approach Using the Vector Space Model	395
<i>Junte Zhang and Jaap Kamps</i>	

XML Mining Track

Overview of the INEX 2008 XML Mining Track: Categorization and Clustering of XML Documents in a Graph of Documents	401
<i>Ludovic Denoyer and Patrick Gallinari</i>	
Semi-supervised Categorization of Wikipedia Collection by Label Expansion	412
<i>Boris Chidlovskii</i>	
Document Clustering with K-tree	420
<i>Christopher M. De Vries and Shlomo Geva</i>	
Using Links to Classify Wikipedia Pages	432
<i>Rianne Kaptein and Jaap Kamps</i>	
Clustering XML Documents Using Frequent Subtrees	436
<i>Sangeetha Kutty, Tien Tran, Richi Nayak, and Yuefeng Li</i>	
UJM at INEX 2008 XML Mining Track	446
<i>Mathias Géry, Christine Largeron, and Christophe Moulin</i>	
Probabilistic Methods for Link-Based Classification at INEX 2008	453
<i>Luis M. de Campos, Juan M. Fernández-Luna, Juan F. Huete, and Alfonso E. Romero</i>	
Utilizing the Structure and Content Information for XML Document Clustering	460
<i>Tien Tran, Sangeetha Kutty, and Richi Nayak</i>	
Self Organizing Maps for the Clustering of Large Sets of Labeled Graphs	469
<i>ShuJia Zhang, Markus Hagenbuchner, Ah Chung Tsoi, and Alessandro Sperduti</i>	
Author Index	483