Lecture Notes in Computer Science 4518

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

Springer

Volume Editors

Norbert Fuhr
Department of Informatics
University of Duisburg-Essen, 47048 Duisburg, Germany
E-mail: 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

Andrew Trotman
Department of Computer Science
University of Otago
Dunedin, New Zealand
E-mail: 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

© 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

Ad Hoc Track

Natural Language Processing Track

Heterogeneous Collection Track

Multimedia Track

Interactive Track

Use Case Track

Document Track