Norbert Fuhr
Mounia Lalmas
Andrew Trotman (Eds.)

LNCS 4518

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 2006
Revised and Selected Papers

Springer

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

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

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

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

Lecture Notes in Computer Science

For information about Vols. 1–4554

please contact your bookseller or Springer

Vol. 4697: L. Choi, Y. Paek, S. Cho (Eds.), Advances in Computer Systems Architecture. XIII, 400 pages. 2007.

Vol. 4682: D.-S. Huang, L. Heutte, M. Loog (Eds.), Advanced Intelligent Computing Theories and Applications. XXVII, 1373 pages. 2007. (Sublibrary LNAI).

Vol. 4681: D.-S. Huang, L. Heutte, M. Loog (Eds.), Advanced Intelligent Computing Theories and Applications. XXVI, 1379 pages. 2007.

Vol. 4673: W.G. Kropatsch, M. Kampel, A. Hanbury (Eds.), Computer Analysis of Images and Patterns. XX, 1006 pages. 2007.

Vol. 4671: V. Malyshkin (Ed.), Parallel Computing Technologies. XIV, 635 pages. 2007.

Vol. 4660: S. Džeroski, J. Todorovski (Eds.), Computational Discovery of Scientific Knowledge. X, 327 pages. 2007. (Sublibrary LNAI).

Vol. 4651: F. Azevedo, P. Barahona, F. Fages, F. Rossi (Eds.), Recent Advances in Constraints. VIII, 185 pages. 2007. (Sublibrary LNAI).

Vol. 4647: R. Martin, M. Sabin, J. Winkler (Eds.), Mathematics of Surfaces XII. IX, 509 pages. 2007.

Vol. 4645: R. Giancarlo, S. Hannenhalli (Eds.), Algorithms in Bioinformatics. XIII, 432 pages. 2007. (Sublibrary LNBI).

Vol. 4643: M.-F. Sagot, M.E.M.T. Walter (Eds.), Advances in Bioinformatics and Computational Biology. XII, 177 pages. 2007. (Sublibrary LNBI).

Vol. 4639: E. Csuhaj-Varj'u, Z. Ésik (Eds.), Fundamentals of Computation Theory. XIV, 508 pages. 2007.

Vol. 4635: B. Kokinov, D.C. Richardson, T.R. Roth-Berghofer, L. Vieu (Eds.), Modeling and Using Context. XIV, 574 pages. 2007. (Sublibrary LNAI).

Vol. 4634: H.R. Nielson, G. Filé (Eds.), Static Analysis. XI, 469 pages. 2007.

Vol. 4633: M. Kamel, A. Campilho (Eds.), Image Analysis and Recognition. XII, 1312 pages. 2007.

Vol. 4632: R. Alhajj, H. Gao, X. Li, J. Li, O.R. Zaïane (Eds.), Advanced Data Mining and Applications. XV, 634 pages. 2007. (Sublibrary LNAI).

Vol. 4628: L.N. de Castro, F.J. Von Zuben, H. Knidel (Eds.), Artificial Immune Systems. XII, 438 pages. 2007.

Vol. 4627: M. Charikar, K. Jansen, O. Reingold, J.D.P. Rolim (Eds.), Approximation, Randomization, and Combinatorial Optimization. XII, 626 pages. 2007.

Vol. 4626: R.O. Weber, M.M. Richter (Eds.), Case-Based Reasoning Research and Development. XIII, 534 pages. 2007. (Sublibrary LNAI).

Vol. 4624: T. Mossakowski, U. Montanari, M. Haveraaen (Eds.), Algebra and Coalgebra in Computer Science. XI, 463 pages. 2007.

Vol. 4622: A. Menezes (Ed.), Advances in Cryptology - CRYPTO 2007. XIV, 631 pages. 2007.

Vol. 4619: F. Dehne, J.-R. Sack, N. Zeh (Eds.), Algorithms and Data Structures. XVI, 662 pages. 2007.

Vol. 4618: S.G. Akl, C.S. Calude, M.J. Dinneen, G. Rozenberg, H.T. Wareham (Eds.), Unconventional Computation. X, 243 pages. 2007.

Vol. 4617: V. Torra, Y. Narukawa, Y. Yoshida (Eds.), Modeling Decisions for Artificial Intelligence. XII, 502 pages. 2007. (Sublibrary LNAI).

Vol. 4616: A. Dress, Y. Xu, B. Zhu (Eds.), Combinatorial Optimization and Applications. XI, 390 pages. 2007.

Vol. 4615: R. de Lemos, C. Gacek, A. Romanovsky (Eds.), Architecting Dependable Systems IV. XIV, 435 pages. 2007.

Vol. 4613: F.P. Preparata, Q. Fang (Eds.), Frontiers in Algorithmics. XI, 348 pages. 2007.

Vol. 4612: I. Miguel, W. Ruml (Eds.), Abstraction, Reformulation, and Approximation. XI, 418 pages. 2007. (Sublibrary LNAI).

Vol. 4611: J. Indulska, J. Ma, L.T. Yang, T. Ungerer, J. Cao (Eds.), Ubiquitous Intelligence and Computing. XXIII, 1257 pages. 2007.

Vol. 4610: B. Xiao, L.T. Yang, J. Ma, C. Muller-Schloer, Y. Hua (Eds.), Autonomic and Trusted Computing. XVIII, 571 pages. 2007.

Vol. 4609: E. Ernst (Ed.), ECOOP 2007 – Object-Oriented Programming. XIII, 625 pages. 2007.

Vol. 4608: H.W. Schmidt, I. Crnkovic, G.T. Heineman, J.A. Stafford (Eds.), Component-Based Software Engineering. XII, 283 pages. 2007.

Vol. 4607: L. Baresi, P. Fraternali, G.-J. Houben (Eds.), Web Engineering. XVI, 576 pages. 2007.

Vol. 4606: A. Pras, M. van Sinderen (Eds.), Dependable and Adaptable Networks and Services. XIV, 149 pages. 2007.

Vol. 4605: D. Papadias, D. Zhang, G. Kollios (Eds.), Advances in Spatial and Temporal Databases. X, 479 pages. 2007.

Vol. 4604: U. Priss, S. Polovina, R. Hill (Eds.), Conceptual Structures: Knowledge Architectures for Smart Applications. XII, 514 pages. 2007. (Sublibrary LNAI).

Vol. 4603: F. Pfenning (Ed.), Automated Deduction – CADE-21. XII, 522 pages. 2007. (Sublibrary LNAI).

Vol. 4602: S. Barker, G.-J. Ahn (Eds.), Data and Applications Security XXI. X, 291 pages. 2007.

Vol. 4600: H. Comon-Lundh, C. Kirchner, H. Kirchner (Eds.), Rewriting, Computation and Proof. XVI, 273 pages. 2007.

Vol. 4599: S. Vassiliadis, M. Berekovic, T.D. Hämäläinen (Eds.), Embedded Computer Systems: Architectures, Modeling, and Simulation. XVIII, 466 pages. 2007.

Vol. 4598: G. Lin (Ed.), Computing and Combinatorics. XII, 570 pages. 2007.

Vol. 4597: P. Perner (Ed.), Advances in Data Mining. XI, 353 pages. 2007. (Sublibrary LNAI).

Vol. 4596: L. Arge, C. Cachin, T. Jurdziński, A. Tarlecki (Eds.), Automata, Languages and Programming. XVII, 953 pages. 2007.

Vol. 4595: D. Bošnački, S. Edelkamp (Eds.), Model Checking Software. X, 285 pages. 2007.

Vol. 4594: R. Bellazzi, A. Abu-Hanna, J. Hunter (Eds.), Artificial Intelligence in Medicine. XVI, 509 pages. 2007. (Sublibrary LNAI).

Vol. 4592: Z. Kedad, N. Lammari, E. Métais, F. Meziane, Y. Rezgui (Eds.), Natural Language Processing and Information Systems. XIV, 442 pages. 2007.

Vol. 4591: J. Davies, J. Gibbons (Eds.), Integrated Formal Methods. IX, 660 pages. 2007.

Vol. 4590: W. Damm, H. Hermanns (Eds.), Computer Aided Verification. XV, 562 pages. 2007.

Vol. 4589: J. Münch, P. Abrahamsson (Eds.), Product-Focused Software Process Improvement. XII, 414 pages. 2007.

Vol. 4588: T. Harju, J. Karhumäki, A. Lepistö (Eds.), Developments in Language Theory. XI, 423 pages. 2007.

Vol. 4587: R. Cooper, J. Kennedy (Eds.), Data Management. XIII, 259 pages. 2007.

Vol. 4586: J. Pieprzyk, H. Ghodosi, E. Dawson (Eds.), Information Security and Privacy. XIV, 476 pages. 2007.

Vol. 4585: M. Kryszkiewicz, J.F. Peters, H. Rybinski, A. Skowron (Eds.), Rough Sets and Intelligent Systems Paradigms. XIX, 836 pages. 2007. (Sublibrary LNAI).

Vol. 4584: N. Karssemeijer, B. Lelieveldt (Eds.), Information Processing in Medical Imaging. XX, 777 pages. 2007.

Vol. 4583: S.R. Della Rocca (Ed.), Typed Lambda Calculi and Applications. X, 397 pages. 2007.

Vol. 4582: J. Lopez, P. Samarati, J.L. Ferrer (Eds.), Public Key Infrastructure. XI, 375 pages. 2007.

Vol. 4581: A. Petrenko, M. Veanes, J. Tretmans, W. Grieskamp (Eds.), Testing of Software and Communicating Systems. XII, 379 pages. 2007.

Vol. 4580: B. Ma, K. Zhang (Eds.), Combinatorial Pattern Matching. XII, 366 pages. 2007.

Vol. 4579: B. M. Hämmerli, R. Sommer (Eds.), Detection of Intrusions and Malware, and Vulnerability Assessment. X, 251 pages. 2007.

Vol. 4578: F. Masulli, S. Mitra, G. Pasi (Eds.), Applications of Fuzzy Sets Theory. XVIII, 693 pages. 2007. (Sublibrary LNAI).

Vol. 4577: N. Sebe, Y. Liu, Y.-t. Zhuang, T.S. Huang (Eds.), Multimedia Content Analysis and Mining. XIII, 513 pages. 2007.

Vol. 4576: D. Leivant, R. de Queiroz (Eds.), Logic, Language, Information and Computation. X, 363 pages. 2007.

Vol. 4575: T. Takagi, T. Okamoto, E. Okamoto, T. Okamoto (Eds.), Pairing-Based Cryptography – Pairing 2007. XI, 408 pages. 2007.

Vol. 4574: J. Derrick, J. Vain (Eds.), Formal Techniques for Networked and Distributed Systems – FORTE 2007. XI, 375 pages. 2007.

Vol. 4573: M. Kauers, M. Kerber, R. Miner, W. Windsteiger (Eds.), Towards Mechanized Mathematical Assistants. XIII, 407 pages. 2007. (Sublibrary LNAI).

Vol. 4572: F. Stajano, C. Meadows, S. Capkun, T. Moore (Eds.), Security and Privacy in Ad-hoc and Sensor Networks. X, 247 pages. 2007.

Vol. 4571: P. Perner (Ed.), Machine Learning and Data Mining in Pattern Recognition. XIV, 913 pages. 2007. (Sublibrary LNAI).

Vol. 4570: H.G. Okuno, M. Ali (Eds.), New Trends in Applied Artificial Intelligence. XXI, 1194 pages. 2007. (Sublibrary LNAI).

Vol. 4569: A. Butz, B. Fisher, A. Krüger, P. Olivier, S. Owada (Eds.), Smart Graphics. IX, 237 pages. 2007.

Vol. 4568: T. Ishida, S. R. Fussell, P. T. J. M. Vossen (Eds.), Intercultural Collaboration. XIII, 395 pages. 2007.

Vol. 4566: M.J. Dainoff (Ed.), Ergonomics and Health Aspects of Work with Computers. XVIII, 390 pages. 2007.

Vol. 4565: D.D. Schmorrow, L.M. Reeves (Eds.), Foundations of Augmented Cognition. XIX, 450 pages. 2007. (Sublibrary LNAI).

Vol. 4564: D. Schuler (Ed.), Online Communities and Social Computing. XVII, 520 pages. 2007.

Vol. 4563: R. Shumaker (Ed.), Virtual Reality. XXII, 762 pages. 2007.

Vol. 4562: D. Harris (Ed.), Engineering Psychology and Cognitive Ergonomics. XXIII, 879 pages. 2007. (Sublibrary LNAI).

Vol. 4561: V.G. Duffy (Ed.), Digital Human Modeling. XXIII, 1068 pages. 2007.

Vol. 4560: N. Aykin (Ed.), Usability and Internationalization, Part II. XVIII, 576 pages. 2007.

Vol. 4559: N. Aykin (Ed.), Usability and Internationalization, Part I. XVIII, 661 pages. 2007.

Vol. 4558: M.J. Smith, G. Salvendy (Eds.), Human Interface and the Management of Information, Part II. XXIII, 1162 pages. 2007.

Vol. 4557: M.J. Smith, G. Salvendy (Eds.), Human Interface and the Management of Information, Part I. XXII, 1030 pages. 2007.

Vol. 4556: C. Stephanidis (Ed.), Universal Access in Human-Computer Interaction, Part III. XXII, 1020 pages. 2007.

Vol. 4555: C. Stephanidis (Ed.), Universal Access in Human-Computer Interaction, Part II. XXII, 1066 pages. 2007.

Table of Contents

Methodology

Overview of INEX 2006 1
Saadia Malik, Andrew Trotman, Mounia Lalmas, and Norbert Fuhr

The Wikipedia XML Corpus 12
Ludovic Denoyer and Patrick Gallinari

INEX 2006 Evaluation Measures 20
Mounia Lalmas, Gabriella Kazai, Jaap Kamps, Jovan Pehcevski, Benjamin Piwowarski, and Stephen Robertson

Choosing an Ideal Recall-Base for the Evaluation of the Focused Task: Sensitivity Analysis of the XCG Evaluation Measures 35
Gabriella Kazai

Ad Hoc Track

A Method of Preferential Unification of Plural Retrieved Elements for XML Retrieval Task 45
Hiroki Tanioka

CISR at INEX 2006 57
Wei Lu, Stephen Robertson, and Andrew Macfarlane

Compact Representations in XML Retrieval 64
Fang Huang, Stuart Watt, David Harper, and Malcolm Clark

CSIRO's Participation in INEX 2006 73
Alexander Krumpholz and David Hawking

Dynamic Element Retrieval in a Semi-structured Collection 82
Carolyn J. Crouch, Donald B. Crouch, Murthy Ganapathibhotla, and Vishal Bakshi

Efficient, Effective and Flexible XML Retrieval Using Summaries 89
M.S. Ali, Mariano Consens, Xin Gu, Yaron Kanza, Flavio Rizzolo, and Raquel Stasiu

Evaluating Structured Information Retrieval and Multimedia Retrieval Using PF/Tijah 104
Thijs Westerveld, Henning Rode, Roel van Os, Djoerd Hiemstra, Georgina Ramírez, Vojkan Mihajlović, and Arjen P. de Vries

EXTIRP: Baseline Retrieval from Wikipedia 115
Miro Lehtonen and Antoine Doucet

Filtering and Clustering XML Retrieval Results 121
Jaap Kamps, Marijn Koolen, and Börkur Sigurbjörnsson

GPX - Gardens Point XML IR at INEX 2006 137
Shlomo Geva

IBM HRL at INEX 06 151
Yosi Mass

Indexing "Reading Paths" for a Structured Information Retrieval at INEX 2006 160
Mathias Géry

Influence Diagrams and Structured Retrieval: Garnata Implementing the SID and CID Models at INEX'06 165
Luis M. de Campos, Juan M. Fernández-Luna, Juan F. Huete, and Alfonso E. Romero

Information Theoretic Retrieval with Structured Queries and Documents 178
Claudio Carpineto, Giovanni Romano, and Caterina Caracciolo

SIRIUS XML IR System at INEX 2006: Approximate Matching of Structure and Textual Content 185
Eugen Popovici, Gildas Ménier, and Pierre-François Marteau

Structured Content-Only Information Retrieval Using Term Proximity and Propagation of Title Terms 200
Michel Beigbeder

Supervised and Semi-supervised Machine Learning Ranking 213
Jean-Noël Vittaut and Patrick Gallinari

The University of Kaiserslautern at INEX 2006 223
Philipp Dopichaj

TopX – AdHoc Track and Feedback Task 233
Martin Theobald, Andreas Broschart, Ralf Schenkel, Silvana Solomon, and Gerhard Weikum

Tuning and Evolving Retrieval Engine by Training on Previous INEX Testbeds 243
Gilles Hubert

Using Language Models and the HITS Algorithm for XML Retrieval ... 253
Benny Kimelfeld, Eitan Kovacs, Yehoshua Sagiv, and Dan Yahav

Using Topic Shifts in XML Retrieval at INEX 2006 261
Elham Ashoori and Mounia Lalmas

XSee: Structure Xposed 271
Roelof van Zwol and Wouter Weerkamp

Natural Language Processing Track

Shallow Parsing of INEX Queries 284
Haïfa Zargayouna, Victor Rosas, and Sylvie Salotti

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

NLPX at INEX 2006 302
Alan Woodley and Shlomo Geva

Heterogeneous Collection Track

The Heterogeneous Collection Track at INEX 2006 312
Ingo Frommholz and Ray Larson

Probabilistic Retrieval Approaches for Thorough and Heterogeneous XML Retrieval 318
Ray R. Larson

Multimedia Track

The INEX 2006 Multimedia Track 331
Thijs Westerveld and Roelof van Zwol

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

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

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

Interactive Track

The Interactive Track at INEX 2006 387
Saadia Malik, Anastasios Tombros, and Birger Larsen

Use Case Track

XML-IR Users and Use Cases 400
Andrew Trotman, Nils Pharo, and Miro Lehtonen

A Taxonomy for XML Retrieval Use Cases 413
Miro Lehtonen, Nils Pharo, and Andrew Trotman

What XML-IR Users May Want 423
Alan Woodley, Shlomo Geva, and Sylvia L. Edwards

Document Track

Report on the XML Mining Track at INEX 2005 and INEX 2006 432
Ludovic Denoyer, Patrick Gallinari, and Anne-Marie Vercoustre

Classifying XML Documents Based on Structure/Content Similarity 444
Guangming Xing, Jinhua Guo, and Zhonghang Xia

Document Mining Using Graph Neural Network 458
S.L. Yong, M. Hagenbuchner, A.C. Tsoi, F. Scarselli, and M. Gori

Evaluating the Performance of XML Document Clustering by Structure Only 473
Tien Tran and Richi Nayak

FAT-CAT: Frequent Attributes Tree Based Classification 485
Jeroen De Knijf

Unsupervised Classification of Text-Centric XML Document Collections 497
Antoine Doucet and Miro Lehtonen

XML Document Mining Using Contextual Self-organizing Maps for Structures 510
M. Kc, M. Hagenbuchner, A.C. Tsoi, F. Scarselli, A. Sperduti, and M. Gori

XML Document Transformation with Conditional Random Fields 525
Rémi Gilleron, Florent Jousse, Isabelle Tellier, and Marc Tommasi

XML Structure Mapping 540
Francis Maes, Ludovic Denoyer, and Patrick Gallinari

Author Index 553

Overview of INEX 2006

Saadia Malik[1], Andrew Trotman[2], Mounia Lalmas[3], and Norbert Fuhr[1]

[1] University of Duisburg-Essen, Duisburg, Germany
{malik,fuhr}@is.informatik.uni-duisburg.de
[2] University of Otago, Dunedin, New Zealand
andrew@cs.otago.ac.nz
[3] Queen Mary, University of London, London, UK
mounia@dcs.qmul.ac.uk

Abstract. Since 2002, INEX has been working towards the goal of establishing an infrastructure, in the form of a large XML test collection and appropriate scoring methods, for the evaluation of content-oriented XML retrieval systems. This paper provides an overview of the work carried out as part of INEX 2006.

1 Introduction

The continuous growth in XML[1] information repositories has been matched by increasing efforts in the development of XML retrieval systems (e.g. [1,2]), in large part aiming at supporting content-oriented XML retrieval. These systems exploit the available structural information, as marked up in XML, in documents, in order to return document components – the so-called XML elements – instead of complete documents in response to a user query. This is of particular benefit for information repositories containing long documents or documents covering a wide variety of topics (e.g. books, user manuals, legal documents), where users' effort to locate relevant content can be reduced by directing them to the most relevant parts of these documents. For example, in response to a user's query on a collection of scientific articles marked-up in XML, an XML retrieval system may return a mixture of paragraph, section, article, or other elements that have been estimated as best answers to the user's query. As the number of XML retrieval systems increases, so does the need to evaluate their effectiveness.

The INitiative for the Evaluation of XML retrieval (INEX)[2], which was set up in 2002, established an infrastructure and provided means, in the form of large test collections and appropriate scoring methods, for evaluating how effective content-oriented XML search systems are. As a result of a collaborative effort during the course of 2006, the INEX test collection has been further extended with the addition of the Wikipedia collection, new topics, and new assessments. Using the constructed test collection and the developed set of measures, the retrieval effectiveness of the participants' XML search engines were evaluated and their results compared.

This paper presents an overview of INEX 2006. Section 2 gives a brief summary of this year's participants. Section 3 provides an overview of the test collection. Section 4

[1] http://www.w3.org/XML/
[2] http://inex.is.informatik.uni-duisburg.de/

N. Fuhr, M. Lalmas, and A. Trotman (Eds.): INEX 2006, LNCS 4518, pp. 1–11, 2007.

outlines the retrieval tasks in the main ad hoc track. Section 5 reports some statistics of the submitted runs. Section 6 describes the relevance assessment phase. The different measures used to evaluate retrieval performance are described in a separate paper [6]. Section 7 provides a short description of the tracks at INEX 2006.

2 Participating Organizations

In reponse to the call for participation, issued in March 2006, 68 organizations registered. Throughout the year a number of groups dropped out due to resource requirements, while 23 groups joined later during the year. The final 50 active groups along with details of their participation is summarized in Table 1.

3 The Test Collection

Test collections consist of three parts: a set of documents, a set of information needs called topics and a set of relevance assessments listing the relevant documents for each topic. Although a test collection for XML retrieval consists of the same three parts, each component is rather different from its traditional information retrieval counterpart.

In traditional information retrieval test collections, documents are considered as units of unstructured text, queries are generally treated as bags of terms or phrases, and relevance assessments provide judgments whether a document as a whole is relevant to a query or not. XML documents organize their content into smaller, nested structural elements. Each of these elements in the document's hierarchy, along with the document itself (the root of the hierarchy), represent a retrievable unit. In addition, with the use of XML query languages, users of an XML retrieval system can express their information need as a combination of content and structural conditions, e.g. users can restrict their search to specific structural elements within the collection. Consequently, the relevance assessments for an XML collection must also consider the structural nature of the documents and provide assessments at different levels of the document hierarchy.

3.1 Documents

INEX 2006 uses a different document collection than in previous years [9], made from English documents from the Wikipedia[3]. The collection is made up of the XML full-texts of 659,388 articles of the Wikipedia project, covering a hierarchy of 113,483 categories, and totaling more than 60 Gigabytes (4.6 Gigabytes without images) and 30 million elements. The collection has a structure containing text, more than 300,000 images and some structured parts corresponding to the Wikipedia templates (about 5000 different tags). The collection has a structure similar to the IEEE collection, which was used up to 2005 in INEX. On average an article contains 161.35 XML nodes, where the average depth of an element is 6.72. For a detailed description of the document collection used for the ad hoc and other tracks at INEX 2006 see [3].

[3] http://en.wikipedia.org

Table 1. List of active INEX 2006 participants

Organizations	Submitted topics	Submitted runs	Assessed topics
Utrecht University	6	11	3
University of California, Berkeley	1	2	3
University of Otago	6	0	3
Queensland University of Technology	6	24	3
Queen Mary University of London	4	12	3
Ecoles des Mines de Saint-Etienne	6	9	3
University of Granada	6	2	3
Indian Statistical Institute	0	2	3
University of Tampere	6	0	3
La Trobe University	6	0	3
University of Kaiserslautern,	6	24	3
City University London	6	13	3
RMIT University	6	12	3
IRIT	9	23	3
Max-Planck-Institut fuer Informatik	6	19	3
University of Cambridge	6	0	3
CSIRO	4	8	3
University of Wollongong in Dubai	5	7	3
University of Amsterdam	8	13	3
Fondazione Ugo Bordoni	6	6	3
The Hebrew University of Jerusalem	6	24	3
Royal School of LIS	6	0	3
University of Toronto	6	1	3
Universität Duisburg-Essen	2	0	1
Oslo University College	3	7	3
University of Waterloo	0	0	3
University of Massachusetts Amherst	6	0	3
Kyungpook National University	0	0	3
University of Rostock	6	3	3
LIP6	5	12	3
CWI and University of Twente	6	23	3
University of Helsinki	4	3	3
The Robert Gordon University	6	6	3
IBM Haifa Research Lab	0	18	3
LIPN	1	0	3
CLIPS-IMAG	6	0	3
Université de Saint-Etienne	6	3	3
Justsystem Corporation	0	12	3
University of South-Brittany	0	20	3
Joint Research Centre	0	0	3
University of Minnesota Duluth	6	14	3
Huazhong University of Science & Technology	0	0	3
Dalhousie University	0	0	3
University College of Boras	0	0	3
Université Libre de Bruxelles	0	0	3
Universidad de Chile			
Organizations participated only in XML document mining track			
INRIA			
Western Kentucky University			
University of Wolongong			
Organization participated only in interactive track			
Rutgers University			

3.2 Topics

Querying XML documents with respect to content can be with or without respect to structure. Taking this into account, INEX identifies two types of topics:

Table 2. Statistics on CO+S topics on the INEX 2006 test collection

	CO+S
No. of topics	125
Average length of title (in words)	4.2
Use of boolean operators (and/or) in title	14
Use of (+/-) in title	61
Use of phrases in title	120
Use of boolean operators (and/or) in castitle	65
Use of (+/-) in castitle	49
Use of phrases in castitle	120
Average length of narrative (in words)	94
Average length of topic description (in words)	14
Average length of topic ontopic_keywords (in words)	6

- Content-only (CO) topics are requests that do not include reference to the document structure. They are, in a sense, the traditional topics used in information retrieval test collections. In XML retrieval, the results to such topics can be elements of various complexity, e.g. at different levels of the XML documents' structure.
- Content-and-structure (CAS) topics are requests that contain conditions referring both to content and structure of a document. These conditions may refer to the content of specific elements (e.g. the elements to be returned must contain a section about a particular topic), or may specify the type of the requested answer elements (e.g. sections should be retrieved).

In previous years a distinction was made between CO and CAS topics. Topic were also designed for use in multiple tracks (such as the natural language track and interactive track) and so contained multiple variant queries for each purpose. Since 2006, these have all been combined into a single topic type: the Content Only + Structure (CO+S) topic. All the information needed by the different tasks and tracks are expressed in each topic, but in different parts of that topic.

Topic Format. Topics are made up of several parts; these parts explain the same information need, but for different purposes.

<narrative>:A detailed explanation of the information need and the description of what makes an element relevant or not. The <narrative> explains not only what information is being sought, but also the context and motivation of the information need, i.e., why the information is being sought and what work-task it might help to solve. Assessments are made on compliance to the <narrative> alone.

<title>: A short explanation of the information need. It serves as a summary of the content of the user's information need. A word in the <title> can have a + or − prefix, where + is used to emphasize an important concept, and − is used to denote an unwanted concept.

<castitle>: A short explanation of the information need, specifying any structural requirements. As with a topic <title>, a word in the <castitle> can have a + or − prefix,