Norbert Fuhr
Mounia Lalmas
Saadia Malik
Zoltán Szlávik (Eds.)

LNCS 3493

Advances in XML Information Retrieval

Third International Workshop of the Initiative for the Evaluation of XML Retrieval, INEX 2004
Dagstuhl Castle, Germany, December 2004
Revised Selected Papers

Springer

Norbert Fuhr Mounia Lalmas
Saadia Malik Zoltán Szlávik (Eds.)

Advances in XML Information Retrieval

Third International Workshop of the Initiative
for the Evaluation of XML Retrieval, INEX 2004
Dagstuhl Castle, Germany, December 6-8, 2004
Revised Selected Papers

Springer

Volume Editors

Norbert Fuhr
University of Duisburg-Essen
Faculty of Engineering Sciences, Information Systems
47048 Duisburg, Germany
E-mail: fuhr@uni-duisburg.de

Mounia Lalmas
Queen Mary University of London, Department of Computer Science
London E1 4NS, England, United Kingdom
E-mail: mounia@dcs.qmul.ac.uk

Saadia Malik
University of Duisburg-Essen
Faculty of Engineering Sciences, Information Systems
47048 Duisburg, Germany
E-mail: malik@is.informatik.uni-duisburg.de

Zoltán Szlávik
Queen Mary University of London, Department of Computer Science
London E1 4NS, England, United Kingdom
E-mail: zolley@dcs.qmul.ac.uk

Library of Congress Control Number: 2005926830

CR Subject Classification (1998): H.3, H.4, H.2

ISSN 0302-9743
ISBN-10 3-540-26166-4 Springer Berlin Heidelberg New York
ISBN-13 978-3-540-26166-7 Springer Berlin Heidelberg New York

Springer is a part of Springer Science+Business Media

springeronline.com

Printed in Germany

Typesetting: Camera-ready by author, data conversion by Scientific Publishing Services, Chennai, India
Printed on acid-free paper SPIN: 11424550 06/3142 5 4 3 2 1 0

Lecture Notes in Computer Science 3493

Preface

The ultimate goal of many information access systems (e.g., digital libraries, the Web, intranets) is to provide the right content to their end-users. This content is increasingly a mixture of text, multimedia, and metadata, and is formatted according to the adopted –W3C standard for information repositories, the so-called eXtensible Markup Language (XML). Whereas 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 campaign (Initiative for the Evaluation of XML Retrieval), which was set up at the beginning of 2002, is to establish infrastructures, XML test suites, and appropriate measurements for evaluating the performance of information retrieval systems that aim at giving effective access to XML content. More precisely, the goal of the INEX initiative is to provide means, in the form of a large XML test collection and appropriate scoring methods, for the evaluation of content-oriented XML retrieval systems.

INEX 2004 was responsible for a range of evaluation activities in the field of XML information retrieval, with five tracks: (1) *Ad Hoc Retrieval Track*, the main track, which can be regarded as a simulation of how a digital library might be used, where a static set of XML documents and their components is searched using a new set of queries (topics) containing both content and structural conditions; (2) *Interactive Track*, which aimed to investigate the behavior of users when interacting with components of XML documents; (3) *Heterogeneous Collection Track*, where retrieval is based on a collection comprising various XML subcollections from different digital libraries, as well as material from other resources; (4) *Relevance Feedback Track*, dealing with relevance feedback methods for XML; and (5) *Natural Language Track*, where natural language formulations of structural conditions of queries have to be answered.

The INEX 2004 workshop, held at Schloss Dagstuhl (Germany), 6–8 December 2004, brought together researchers in the field of XML retrieval who participated in the INEX 2004 evaluation campaign. Participants were able to present and discuss their approaches to XML retrieval. These proceedings contain revised papers describing work carried out during INEX 2004 in the various tracks by the participants.

INEX is partly funded by the DELOS Network of Excellence on Digital Libraries, to which we are very thankful. We would also like to thank the IEEE Computer Society for providing us the XML document collection. Special thanks go to Shlomo Geva for setting up the WIKI server and Gabriella Kazai for helping with the various documentation. We gratefully acknowledge the involvement of Börkur Sigurbjörnsson and Andrew Trotman (topic format specification), Benjamin Piwowarski (online assessment tool), and Gabriella Kazai and Arjen de Vries (metrics). The organizers of the various tracks did a great job and their work is greatly appreciated: Anastasios Tombros, Birger Larsen, Thomas Rölleke, Carolyn Crouch, Shlomo Geva and Tony Sahama. Finally, we would like to thank the participating organizations and people for their participation in INEX 2004.

March 2005

Norbert Fuhr
Mounia Lalmas
Saadia Malik
Zoltán Szlávik

Organizers

Organizers

Project Leaders

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

Contact Person

Saadia Malik, University of Duisburg-Essen

Topic Format Specification

Börkur Sigurbjörnsson, University of Amsterdam
Andrew Trotman, University of Otago

Online Relevance Assessment Tool

Benjamin Piwowarski, University of Chile

Metrics

Gabriella Kazai, Queen Mary University of London
Arjen P. de Vries, Centre for Mathematics and Computer Science

Interactive Track

Birger Larsen, Royal School of Library and Information Science
Saadia Malik, University of Duisburg-Essen
Anastasios Tombros, Queen Mary University of London

Relevance Feedback Track

Carolyn Crouch, University of Minnesota-Duluth
Mounia Lalmas, Queen Mary University of London

Heterogeneous Collection Track

Thomas Rölleke, Queen Mary University of London
Zoltán Szlávik, Queen Mary University of London

Natural Language Processing

Shlomo Geva, Queensland University of Technology
Tony Sahama, Queensland University of Technology

Lecture Notes in Computer Science

For information about Vols. 1–3411

please contact your bookseller or Springer

Vol. 3532: A. Gómez-Pérez, J. Euzenat (Eds.), The Semantic Web: Research and Applications. XV, 728 pages. 2005.

Vol. 3526: S.B. Cooper, B. Löwe, L. Torenvliet (Eds.), New Computational Paradigms. XVII, 574 pages. 2005.

Vol. 3525: A.E. Abdallah, C.B. Jones, J.W. Sanders (Eds.), Communicating Sequential Processes. XIV, 321 pages. 2005.

Vol. 3518: T.B. Ho, D. Cheung, H. Li (Eds.), Advances in Knowledge Discovery and Data Mining. XXI, 864 pages. 2005. (Subseries LNAI).

Vol. 3517: H.S. Baird, D.P. Lopresti (Eds.), Human Interactive Proofs. IX, 143 pages. 2005.

Vol. 3516: V.S. Sunderam, G.D.v. Albada, P.M.A. Sloot, J.J. Dongarra (Eds.), Computational Science – ICCS 2005, Part III. LXIII, 1143 pages. 2005.

Vol. 3515: V.S. Sunderam, G.D.v. Albada, P.M.A. Sloot, J.J. Dongarra (Eds.), Computational Science – ICCS 2005, Part II. LXIII, 1101 pages. 2005.

Vol. 3514: V.S. Sunderam, G.D.v. Albada, P.M.A. Sloot, J.J. Dongarra (Eds.), Computational Science – ICCS 2005, Part I. LXIII, 1089 pages. 2005.

Vol. 3513: A. Montoyo, R. Mu\˜noz, E. Métais (Eds.), Natural Language Processing and Information Systems. XII, 408 pages. 2005.

Vol. 3510: T. Braun, G. Carle, Y. Koucheryavy, V. Tsaoussidis (Eds.), Wired/Wireless Internet Communications. XIV, 366 pages. 2005.

Vol. 3508: P. Bresciani, P. Giorgini, B. Henderson-Sellers, G. Low, M. Winikoff (Eds.), Agent-Oriented Information Systems II. X, 227 pages. 2005. (Subseries LNAI).

Vol. 3505: V. Gorodetsky, J. Liu, V. A. Skormin (Eds.), Autonomous Intelligent Systems: Agents and Data Mining. XIII, 303 pages. 2005. (Subseries LNAI).

Vol. 3503: S.E. Nikoletseas (Ed.), Experimental and Efficient Algorithms. XV, 624 pages. 2005.

Vol. 3502: F. Khendek, R. Dssouli (Eds.), Testing of Communicating Systems. X, 381 pages. 2005.

Vol. 3501: B. Kégl, G. Lapalme (Eds.), Advances in Artificial Intelligence. XV, 458 pages. 2005. (Subseries LNAI).

Vol. 3500: S. Miyano, J. Mesirov, S. Kasif, S. Istrail, P. Pevzner, M. Waterman (Eds.), Research in Computational Molecular Biology. XVII, 632 pages. 2005. (Subseries LNBI).

Vol. 3499: A. Pelc, M. Raynal (Eds.), Structural Information and Communication Complexity. X, 323 pages. 2005.

Vol. 3498: J. Wang, X. Liao, Z. Yi (Eds.), Advances in Neural Networks – ISNN 2005, Part III. L, 1077 pages. 2005.

Vol. 3497: J. Wang, X. Liao, Z. Yi (Eds.), Advances in Neural Networks – ISNN 2005, Part II. L, 947 pages. 2005.

Vol. 3496: J. Wang, X. Liao, Z. Yi (Eds.), Advances in Neural Networks – ISNN 2005, Part II. L, 1055 pages. 2005.

Vol. 3495: P. Kantor, G. Muresan, F. Roberts, D.D. Zeng, F.-Y. Wang, H. Chen, R.C. Merkle (Eds.), Intelligence and Security Informatics. XVIII, 674 pages. 2005.

Vol. 3494: R. Cramer (Ed.), Advances in Cryptology – EUROCRYPT 2005. XIV, 576 pages. 2005.

Vol. 3493: N. Fuhr, M. Lalmas, S. Malik, Z. Szlávik (Eds.), Advances in XML Information Retrieval. XI, 438 pages. 2005.

Vol. 3492: P. Blache, E. Stabler, J. Busquets, R. Moot (Eds.), Logical Aspects of Computational Linguistics. X, 363 pages. 2005. (Subseries LNAI).

Vol. 3489: G.T. Heineman, I. Crnkovic, H.W. Schmidt, J.A. Stafford, C. Szyperski, K. Wallnau (Eds.), Component-Based Software Engineering. XI, 358 pages. 2005.

Vol. 3488: M.-S. Hacid, N.V. Murray, Z.W. Raś, S. Tsumoto (Eds.), Foundations of Intelligent Systems. XIII, 700 pages. 2005. (Subseries LNAI).

Vol. 3486: T. Helleseth, D. Sarwate, H.-Y. Song, K. Yang (Eds.), Sequences and Their Applications - SETA 2004. XII, 451 pages. 2005.

Vol. 3483: O. Gervasi, M.L. Gavrilova, V. Kumar, A. Laganà, H.P. Lee, Y. Mun, D. Taniar, C.J.K. Tan (Eds.), Computational Science and Its Applications – ICCSA 2005, Part IV. XXVII, 1362 pages. 2005.

Vol. 3482: O. Gervasi, M.L. Gavrilova, V. Kumar, A. Laganà, H.P. Lee, Y. Mun, D. Taniar, C.J.K. Tan (Eds.), Computational Science and Its Applications – ICCSA 2005, Part III. LXVI, 1340 pages. 2005.

Vol. 3481: O. Gervasi, M.L. Gavrilova, V. Kumar, A. Laganà, H.P. Lee, Y. Mun, D. Taniar, C.J.K. Tan (Eds.), Computational Science and Its Applications – ICCSA 2005, Part II. LXIV, 1316 pages. 2005.

Vol. 3480: O. Gervasi, M.L. Gavrilova, V. Kumar, A. Laganà, H.P. Lee, Y. Mun, D. Taniar, C.J.K. Tan (Eds.), Computational Science and Its Applications – ICCSA 2005, Part I. LXV, 1234 pages. 2005.

Vol. 3479: T. Strang, C. Linnhoff-Popien (Eds.), Location- and Context-Awareness. XII, 378 pages. 2005.

Vol. 3478: C. Jermann, A. Neumaier, D. Sam (Eds.), Global Optimization and Constraint Satisfaction. XIII, 193 pages. 2005.

Vol. 3477: P. Herrmann, V. Issarny, S. Shiu (Eds.), Trust Management. XII, 426 pages. 2005.

Vol. 3475: N. Guelfi (Ed.), Rapid Integration of Software Engineering Techniques. X, 145 pages. 2005.

Vol. 3468: H.W. Gellersen, R. Want, A. Schmidt (Eds.), Pervasive Computing. XIII, 347 pages. 2005.

Vol. 3467: J. Giesl (Ed.), Term Rewriting and Applications. XIII, 517 pages. 2005.

Vol. 3465: M. Bernardo, A. Bogliolo (Eds.), Formal Methods for Mobile Computing. VII, 271 pages. 2005.

Vol. 3464: S.A. Brueckner, G.D.M. Serugendo, A. Karageorgos, R. Nagpal (Eds.), Engineering Self-Organising Systems. XIII, 299 pages. 2005. (Subseries LNAI).

Vol. 3463: M. Dal Cin, M. Kaâniche, A. Pataricza (Eds.), Dependable Computing - EDCC 2005. XVI, 472 pages. 2005.

Vol. 3462: R. Boutaba, K. Almeroth, R. Puigjaner, S. Shen, J.P. Black (Eds.), NETWORKING 2005. XXX, 1483 pages. 2005.

Vol. 3461: P. Urzyczyn (Ed.), Typed Lambda Calculi and Applications. XI, 433 pages. 2005.

Vol. 3460: Ö. Babaoglu, M. Jelasity, A. Montresor, C. Fetzer, S. Leonardi, A. van Moorsel, M. van Steen (Eds.), Self-star Properties in Complex Information Systems. IX, 447 pages. 2005.

Vol. 3459: R. Kimmel, N.A. Sochen, J. Weickert (Eds.), Scale Space and PDE Methods in Computer Vision. XI, 634 pages. 2005.

Vol. 3458: P. Herrero, M.S. Pérez, V. Robles (Eds.), Scientific Applications of Grid Computing. X, 208 pages. 2005.

Vol. 3456: H. Rust, Operational Semantics for Timed Systems. XII, 223 pages. 2005.

Vol. 3455: H. Treharne, S. King, M. Henson, S. Schneider (Eds.), ZB 2005: Formal Specification and Development in Z and B. XV, 493 pages. 2005.

Vol. 3454: J.-M. Jacquet, G.P. Picco (Eds.), Coordination Models and Languages. X, 299 pages. 2005.

Vol. 3453: L. Zhou, B.C. Ooi, X. Meng (Eds.), Database Systems for Advanced Applications. XXVII, 929 pages. 2005.

Vol. 3452: F. Baader, A. Voronkov (Eds.), Logic for Programming, Artificial Intelligence, and Reasoning. XI, 562 pages. 2005. (Subseries LNAI).

Vol. 3450: D. Hutter, M. Ullmann (Eds.), Security in Pervasive Computing. XI, 239 pages. 2005.

Vol. 3449: F. Rothlauf, J. Branke, S. Cagnoni, D.W. Corne, R. Drechsler, Y. Jin, P. Machado, E. Marchiori, J. Romero, G.D. Smith, G. Squillero (Eds.), Applications of Evolutionary Computing. XX, 631 pages. 2005.

Vol. 3448: G.R. Raidl, J. Gottlieb (Eds.), Evolutionary Computation in Combinatorial Optimization. XI, 271 pages. 2005.

Vol. 3447: M. Keijzer, A. Tettamanzi, P. Collet, J.v. Hemert, M. Tomassini (Eds.), Genetic Programming. XIII, 382 pages. 2005.

Vol. 3444: M. Sagiv (Ed.), Programming Languages and Systems. XIII, 439 pages. 2005.

Vol. 3443: R. Bodik (Ed.), Compiler Construction. XI, 305 pages. 2005.

Vol. 3442: M. Cerioli (Ed.), Fundamental Approaches to Software Engineering. XIII, 373 pages. 2005.

Vol. 3441: V. Sassone (Ed.), Foundations of Software Science and Computational Structures. XVIII, 521 pages. 2005.

Vol. 3440: N. Halbwachs, L.D. Zuck (Eds.), Tools and Algorithms for the Construction and Analysis of Systems. XVII, 588 pages. 2005.

Vol. 3439: R.H. Deng, F. Bao, H. Pang, J. Zhou (Eds.), Information Security Practice and Experience. XII, 424 pages. 2005.

Vol. 3438: H. Christiansen, P.R. Skadhauge, J. Villadsen (Eds.), Constraint Solving and Language Processing. VIII, 205 pages. 2005. (Subseries LNAI).

Vol. 3437: T. Gschwind, C. Mascolo (Eds.), Software Engineering and Middleware. X, 245 pages. 2005.

Vol. 3436: B. Bouyssounouse, J. Sifakis (Eds.), Embedded Systems Design. XV, 492 pages. 2005.

Vol. 3434: L. Brun, M. Vento (Eds.), Graph-Based Representations in Pattern Recognition. XII, 384 pages. 2005.

Vol. 3433: S. Bhalla (Ed.), Databases in Networked Information Systems. VII, 319 pages. 2005.

Vol. 3432: M. Beigl, P. Lukowicz (Eds.), Systems Aspects in Organic and Pervasive Computing - ARCS 2005. X, 265 pages. 2005.

Vol. 3431: C. Dovrolis (Ed.), Passive and Active Network Measurement. XII, 374 pages. 2005.

Vol. 3429: E. Andres, G. Damiand, P. Lienhardt (Eds.), Discrete Geometry for Computer Imagery. X, 428 pages. 2005.

Vol. 3428: Y.-J. Kwon, A. Bouju, C. Claramunt (Eds.), Web and Wireless Geographical Information Systems. XII, 255 pages. 2005.

Vol. 3427: G. Kotsis, O. Spaniol (Eds.), Wireless Systems and Mobility in Next Generation Internet. VIII, 249 pages. 2005.

Vol. 3423: J.L. Fiadeiro, P.D. Mosses, F. Orejas (Eds.), Recent Trends in Algebraic Development Techniques. VIII, 271 pages. 2005.

Vol. 3422: R.T. Mittermeir (Ed.), From Computer Literacy to Informatics Fundamentals. X, 203 pages. 2005.

Vol. 3421: P. Lorenz, P. Dini (Eds.), Networking - ICN 2005, Part II. XXXV, 1153 pages. 2005.

Vol. 3420: P. Lorenz, P. Dini (Eds.), Networking - ICN 2005, Part I. XXXV, 933 pages. 2005.

Vol. 3419: B. Faltings, A. Petcu, F. Fages, F. Rossi (Eds.), Constraint Satisfaction and Constraint Logic Programming. X, 217 pages. 2005. (Subseries LNAI).

Vol. 3418: U. Brandes, T. Erlebach (Eds.), Network Analysis. XII, 471 pages. 2005.

Vol. 3416: M. Böhlen, J. Gamper, W. Polasek, M.A. Wimmer (Eds.), E-Government: Towards Electronic Democracy. XIII, 311 pages. 2005. (Subseries LNAI).

Vol. 3415: P. Davidsson, B. Logan, K. Takadama (Eds.), Multi-Agent and Multi-Agent-Based Simulation. X, 265 pages. 2005. (Subseries LNAI).

Vol. 3414: M. Morari, L. Thiele (Eds.), Hybrid Systems: Computation and Control. XII, 684 pages. 2005.

Vol. 3412: X. Franch, D. Port (Eds.), COTS-Based Software Systems. XVI, 312 pages. 2005.

Table of Contents

Methodology

Ad Hoc Retrieval

Ad Hoc Retrieval and Relevance Feedback

Relevance Feedback

Ad Hoc Retrieval and Heterogeneous Document Collection

Heterogeneous Document Collection

Natural Language Processing of Topics

Interactive Studies

Overview of INEX 2004

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

[1] Information Systems, University of Duisburg-Essen, Duisburg, Germany
malik@is.informatik.uni-duisburg.de
[2] Department of Computer Science, Queen Mary University of London, London, UK
mounia@dcs.qmul.ac.uk
[3] Information Systems, University of Duisburg-Essen, Duisburg, Germany
fuhr@uni-duisburg.de

1 Introduction

The widespread use of the eXtensible Markup Language (XML) in scientific data repositories, digital libraries and on the web, brought about an explosion in the development of XML retrieval systems. These systems exploit the logical structure of documents, which is explicitly represented by the XML markup: instead of whole documents, only components thereof (the so-called XML elements) are retrieved in response to a user query. This means that an XML retrieval system needs not only to find relevant information in the XML documents, but also determine the appropriate level of granularity to return to the user, and this with respect to both content and structural conditions.

Evaluating the effectiveness of XML retrieval systems requires a test collection (XML documents, tasks/topics, and relevance judgements) where the relevance assessments are provided according to a relevance criterion that takes into account the imposed structural aspects. A test collection as such has been built as a result of three rounds of the Initiative for the Evaluation of XML Retrieval[1] (INEX 2002, INEX 2003 and INEX 2004). The aim of this initiative is to provide means, in the form of large testbeds and appropriate scoring methods, for the evaluation of content-oriented retrieval of XML documents.

This paper presents an overview of INEX 2004. In section 2, we give a brief summary of the participants. Section 3 provides an overview of the test collection along with the description of how the collection was constructed. Section 4 outlines the retrieval tasks in the main track, which is concerned with the ad hoc retrieval of XML documents. Section 5 briefly reports on the submission runs for the retrieval tasks, and Section 6 describes the relevance assessment phase. The different metrics used are discussed in Section 7, followed by a summary of the evaluation results in Section 8. Section 9 presents a short description of four new tracks that started in INEX 2004, namely the heterogenous collection track, the relevance feedback track, the natural language processing track and the interactive track. The paper finishes with some conclusions and an outlook for INEX 2005.

[1] http://inex.is.informatik.uni-duisburg.de/

N. Fuhr et al. (Eds.): INEX 2004, LNCS 3493, pp. 1–15, 2005.

2 Participating Organisations

In response to the call for participation issued in March 2004, around 55 organisations registered from 20 different countries within six weeks. Throughout the year, the number of participants decreased due to insufficient contribution while a number of new groups joined later at the assessment phase. The active participants are listed in Table 1.

3 The Test Collection

The INEX test collection, as for any IR test collection aiming at evaluating retrieval effectiveness, is composed of three parts: the set of documents, the set of topics, and the relevance assessments (these are described in Section 6).

3.1 Documents

The document collection was donated by the IEEE Computer Society. It consists of the full-text of 12,107 articles, marked up in XML, from 12 magazines and 6 transactions of the IEEE Computer Society's publications, covering the period of 1995-2002, and totalling 494 MB in size, and 8 millions in number of elements. The collection contains scientific articles of varying length. On average, an article contains 1,532 XML nodes, where the average depth of the node is 6.9. More details can be found in [3].

3.2 Topics

As in previous years, in INEX 2004 we distinguish two types of topics, reflecting two user profiles, where the users differ in the amount of knowledge they have about the structure of the collection:

- **Content-only (CO) topics** are requests that ignore the document structure and contain only content related conditions, e.g. only specify what a document/component should be about (without specifying what that component is). The CO topics simulate users who do not (want to) know, or do not want to use, the actual structure of the XML documents. This profile is likely to fit most users searching XML digital libraries.
- **Content-and-structure (CAS) topics** are topic statements that contain explicit references to the XML structure, and explicitly specify the contexts of the user's interest (e.g. target elements) and/or the contexts of certain search concepts (e.g. containment conditions). The CAS topics simulate users that have some knowledge of the structure of the XML. Those users might want to use this knowledge to try to make their topics more concrete, by adding structural constraints. This user profile could fit librarians that have some knowledge of the collection structure.

The topic format and guidelines were based on TREC guidelines, but were modified to accommodate the two types of topics used in INEX. Both CO and CAS topics are made up of four parts. The parts explain the same information need, but for different purposes.

Table 1. List of active INEX 2004 participants

Organisations	Assessed topics	no of runs submitted
University of Amsterdam	2	6
University of California, Berkeley	2	5
VSB-Technical University of Ostrava	2	0
RMIT University	1	6
University of Otago	2	1
IBM Haifa Research Lab	2	6
University of Illinois at Urbana-Champaign	2	0
Nara Institute of Science and Technology	2	4
University of Wollongong in Dubai	2	0
Fondazione Ugo Bordoni	2	3
IRIT	2	6
Ecoles des Mines de Saint-Etienne, France	1	2
University of Munich (LMU)	2	5
Queen Mary University of London	2	0
Royal School of LIS	1	3
LIP6	1	6
University of Tampere	2	6
University of Helsinki	2	3
Carnegie Mellon University	1	6
Cirquid project (CWI and University of Twente)	2	6
The Selim and Rachel Benin School of Engineering and Computer Science	2	0
University of Minnesota Duluth	1	2
Bamberg University	2	0
UCLA	2	6
Max-Planck-Institut fuer Informatik	4	4
Kyungpook National University	2	4
Utrecht University	1	6
The Robert Gordon University	0	2
University of Milano	2	0
Oslo University College	2	6
Cornell University	2	0
Universität Rostock	2	0
Universidade Estadual de Montos Claros	2	6
INRIA	2	0
LIMSI/CNRS	2	0
University of Waterloo	2	3
Queensland University of Technology	2	6
Indiana University	2	2
University of Granada	2	0
University of Kaiserslautern	2	0
The University of Iowa	2	0
Rutgers University	2	0
IIT Information Retrieval Lab	2	0

- **Title:** a short explanation of the information need. It serves as a summary of both the content and, in the case of CAS topics, also the structural requirements of the user's information need. For the expression of these constraints the Narrowed Extended XPath I (NEXI) query syntax is used [8].
- **Description:** a one or two sentence natural language definition of the information need.
- **Narrative:** a detailed explanation of the information need and the description of what makes a document/component relevant or not. The narrative was there to explain not only what information is being sought for, 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. The latter was required for the interactive track (see Section 9.4).
- **Keywords:** a set of comma-separated scan terms that were used in the collection exploration phase of the topic development process (see later) to retrieve relevant documents/ components. Scan terms may be single words or phrases and may include synonyms, and terms that are broader or narrower terms than those listed in the topic description or title.

The title and the description must be interchangeable, which was required for the natural language processing track (see Section 9.3). The DTD of the topics is shown in Figure 1.

```
<!ELEMENT inex_topic  (title,description,narrative,keywords)>
<!ATTLIST inex_topic
  topic_id    CDATA  #REQUIRED
  query_type  CDATA  #REQUIRED
  ct_no       CDATA  #REQUIRED
>
<!ELEMENT title (#PCDATA)>
<!ELEMENT description   (#PCDATA)>
<!ELEMENT narrative     (#PCDATA)>
<!ELEMENT keywords      (#PCDATA)>
```

Fig. 1. Topic DTD

The attributes of the topic are: topic_id (which ranges from 127 to 201), query_type (with value CAS or CO) and ct_no, which refers to the candidate topic number (which ranges from 1 to 198). Examples of both types of topics can be seen in Figure 2 and Figure 3.

The topics were created by participating groups. Each participant was asked to submit up to 6 candidate topics (3 CO and 3 CAS). A detailed guideline was provided to the participants for the topic creation. Four steps were identified for this process: 1) Initial Topic Statement creation 2) Collection Exploration 3) Topic Refinement and 4) Topic Selection. The first three steps were performed by the participants themselves while the selection of topics was decided by the organisers.

During the first step, participants created their initial topic statement. These were treated as a user's description of his/her information need and were formed without

```
<inex_topic topic_id="127" query_type="CAS" ct_no="13">
<title>//sec//(p| fgc)[about( ., Godel Lukasiewicz and other
fuzzy implication definitions)]</title>
<description>Find paragraphs or figure-captions containing the
definition of Godel, Lukasiewicz or other fuzzy-logic
implications</description>
<narrative>Any relevant element of a section must contain the
definition of a fuzzy-logic implication operator or a pointer to
the element of the same article where the definition can be
found. Elements containing criteria for identifying or comparing
fuzzy implications are also of interest. Elements which discuss
or introduce non-implication fuzzy operators are not relevant.
</narrative>
<keywords>Godel implication, Lukasiewicz implication, fuzzy
implications, fuzzy-logic implication </keywords>
</inex_topic>
```

Fig. 2. A CAS topic from the INEX 2004 test collection

```
<inex_topic topic_id="162" query_type="CO" ct_no="1">
<title> Text and Index Compression Algorithms </title>
<description>Any type of coding algorithm for text and index
compression</description>
<narrative>We have developed an information retrieval system
implementing compression techniques for indexing documents. We
are interested in improving the compression rate of the system
preserving a fast access and decoding of the data. A relevant
document/component should introduce new algorithms or compares
the performance of existing text-coding techniques for text and
index compression. A document/component discussing the cost of
text compression for text coding and decoding is highly relevant.
Strategies for dictionary compression are not
relevant.</narrative>
<keywords>text compression, text coding, index compression
algorithm</keywords>
</inex_topic>
```

Fig. 3. A CO topic from the INEX 2004 test collection

regard to system capabilities or collection peculiarities to avoid artificial or collection biased queries. During the collection exploration phase, participants estimated the number of relevant documents/components to their candidate topics. The HyREX retrieval system [1] was provided to participants to perform this task. Participants had to judge the top retrieved results and were asked to record the relevant document/component (XPath) paths in the top 25 retrieved components/documents. Those topics having at least 2 relevant documents/components and less than 20 documents/components in the top 25 retrieved elements could be submitted as candidate topics. In the topic refinement stage, the topics were finalised ensuring coherency and that each part of the topic could be used in stand-alone fashion.