Task 7.3 - Initiative for the Evaluation of XML Access & Retrieval (INEX)

Mounia Lalmas & Tassos Tombros

Queen Mary University of London

Objectives of INEX

Promote research and stimulate development of XML information access and retrieval, through

- Creation of evaluation infrastructure and organisation of regular evaluation campaigns for system testing
- Building of an XML information access and retrieval research community
- Construction of test-suites

INEX 2005 has allowed a new community in XML information access to emerge, as shown by the number of publications (64 - not final- in 2005, 37 in 2004 and 13 in 2003). INEX is allowing research in XML information access to thrive.

INEX Evaluation Initiative: Background

- Sponsored by DELOS Network of Excellence for Digital Libraries under FP6 – IST programme
- Mainly dependent on voluntary efforts
- Coordination is distributed for tasks and tracks

Main Institutions involved in Coordination for 2005

- University of Amsterdam, NL
- University of Otago, NZ
- University of Chile, CL
- CWI, NL
- Carnegie Mellon University, USA
- IBM Research Lab, IL
- University of Minnesota Duluth, USA
- University of Paris 6, FR

- Queensland University of Technology, AUS
- University of California, Berkeley, USA
- Royal School of LIS, DK
- Queen Mary University of London, UK
- University of Duisburg-Essen, DE
- INRIA-Rocquencourt, FR
- Utrecht University, NL

INEX 2005: Evaluation Tracks

INEX 2005 offered eight tracks designed to evaluate the performance of XML access systems:

- XML retrieval (ad-hoc)
- Interactive track (iTrack)
- Heterogeneous track
- Relevance feedback track
- Natural language query processing track
- Multimedia XML retrieval track **(New)**
- Document mining track (in collaboration with PASCAL network) **(New)**

Max-Planck-Institut fuer Informatik, Germany
Information Studies, Royal School of LIS, Denmark
University of California, Berkeley, USA
Peking University, China
University of Granada, Spain
University of Amsterdam, The Netherlands
University of Otago, New Zealand
Queen Mary University of London, UK
University of Toronto, Canada
Utrecht University, The Netherlands
City University London, UK
University of Kaiserslautern, Germany
INRIA-Rocquencourt, France
University of Wollongong in Dubai
IRIT - Toulouse, France
RMIT University, Australia
Ecoles des Mines de Saint-Etienne, France
Queensland University of Technology, Australia
Universtity of Klagenfurt, Austria
Fondazione Ugo Bordoni, Italy
University of Tampere, Finland
Carnegie Mellon University, USA
Cornell University, USA
University of Illinois at Urbana-Champaign, USA
IBM Haifa Research Lab, Israel
Ochanomizu University, Japan
The Hebrew University of Jerusalem, Israel
Laboratoire d'Informatique de Paris 6, France
University of Minnesota Duluth, USA
University of Rostock, Germany
University of California, Los Angeles, USA
University of Udine, Italy

University of Rostock, Germany
University of California, Los Angeles, USA

University of Udine, Italy
University of South-Brittany, France
Nagoya University, Japan
University of Waterloo, Canada
Rutgers University, USA
Kyungpook National University, Korea
University of Chile, Chile
Hiroshima City University, Japan
University of Helsinki, Finland
AT&T Labs-Research, USA
Microsoft Research Lab Cambridge, UK
University of Twente, The Netherlands
Centre for Mathematics & Computer Science (CWI), NL
University of Utah, USA
University Duisburg-Essen, Germany
University of Ostrava, Czech Republic
Hong Kong Baptist University, Hong Kong
University of Sheffield, UK
Oslo University College, Norway
L3S Research Center, Germany
University of Michigan, USA
CLIPS-IMAG Grenoble, France
Wuhan University, China
Nara Institute of Science and Technology, Japan
Ritsumeikan University, Japan
University of Tsukuba, Japan
State University of Montes Claros, Montes Claros(MG), Brazil
INRIA Sophia Antipolis
Charles de Gaulle University - Lille 3
University of Siena, Italy
Australian Research Council, Canberra, Australia
University of Wollongong, Wollongong, Australia

University of Padova, Italy

64 participants: 32 Europe; 12 N.America; 10 Asia; 5 Oceania; 5 other

INEX 2005 Results: Expansion of Test Collections

- The INEX corpus until 2004 was composed of full-texts, marked up in XML, of 12,107 articles of the **IEEE Computer Society's** publications from 12 magazines and 6 transactions, covering the period of 1995-2002, and totalling 494 MB. An addition of 4,712 articles totalling 241 MB from 2002-2004 have been added to the collection, thus reaching a total of 16,819 articles.

- A new corpus was acquired for the multimedia track in 2005, made available by the **Lonely Planet** organisation. It consists of 462 XML documents with information about destinations, that is particularly useful for travellers that want to find interesting details for their next holiday or business trip. This particular collection is referred to as the "WorldGuide" and can be viewed online at: http://www.lonelyplanet.com/worldguide.

INEX 2005 Results: Expansion of Test Collections

- For the heterogeneous track the following document collections have been made available:
  - Berkeley (Library catalog entries for CS literature): 12,800 XML items
  - CompuScience (Bibliographic entries from the Computer Science database of FIZ Karlsruhe): 250,987 XML items.
  - bibdbpub (BibTeX converted to XML by the IS group at Univ. of Duisburg-Essen): 3465 XML items.
  - dblp (Bibliographic entries from the Digital Bibliography & Library Project in Trier): 501,101 XML items.
  - hcibib (Human-Computer Interaction Resources, bibliography from www.hcibib.org): 26,402 XML items.
  - qmuldcsdpub (Publications database of QMUL Department of Computer Science): 2024 XML items.
  - ZDNet (Articles and Comments) provided by ZDNet.com : 96,351 items (4734 Articles and 91,617 comments on those articles) (new in 2005)
- For the document mining track, 2 new collections were developed:
  - WIPO corpus is composed of 75,250 XML documents
  - MovieDB corpus (based on the Internet Movie Database) consists of 9463 XML documents

- The **interactive track** expanded in the numbers of tasks offered and in the number of participating groups; the track tries to answer some fundamental questions of XML IR

- The **heterogeneous track** expanded by studying new collections with different DTDs and their effect on XML IR system effectiveness

- The **relevance feedback** track investigated approaches for queries that also include structural hints (rather than content-only queries in 2004)

- The **natural language query processing track** included a new task that allows new participants with NLP expertise to join INEX without the need to develop a search engine, and thus encouraging wider accessibility

INEX 2005: Results Research in New Directions II

- The aim of the **document mining track**, done in collaboration with the PASCAL network of Excellence (http://www.pascal-network.org/), is to develop machine learning methods for structured data mining and to evaluate these methods for XML document mining tasks. The track in 2005 focused on classification and clustering for XML documents

- The main objective of the **multimedia track** is to provide an evaluation platform/forum for structured document access systems that do not only include text in the retrieval process, but also other types of media, such as images, speech, and video

INEX 2005: Results
Research in New Directions III

- Main retrieval task (ad-hoc track), a simulation of how a XML content-based library might be used

- 3 sub-tasks depending on structural constraints
  - **Content-Only (CO)** sub-task, queries ignore document structure and contain only content-related conditions
  - Extension of the CO sub-task that includes structural hints (+S sub-task), where user may decide to add structural hints to query to narrow down number of returned documents resulting from a CO query
  - **Content and Structure (CAS)** sub-task, structural constraints explicitly stated:
    - where to look for the relevant elements (i.e. support elements)
    - what type of elements to return (i.e. target elements)
    - strict and vague interpretations applied to both support and target elements, total of four strategies

INEX 2005: Results
Research in New Directions IV

- For ad hoc track, assumptions on what users want as output of an XML retrieval system led to three different strategies:
  - In a **focussed strategy**, we assume that a user prefers a single element that most exhaustively discusses the topic of the query (most exhaustive element), while at the same time it is most specific only to that topic (most specific element)
  - In a **thorough strategy**, we assume that a user prefers all highly exhaustive and specific elements
  - In a **fetch & browse strategy** we assume that a user is interested in highly exhaustive and specific elements that are contained only within highly relevant articles

INEX 2005: Results
Research in New Directions V

- INEX adopted a new set of metrics, the eXtended Cumulated Gain (XCG) metrics to support the evaluation of XML access systems
  - evaluation framework that considers dependency among XML document components
  - system and user-oriented evaluation aspects and recall & precision-like qualities are measured.
    - User-oriented measures allow to reason about a system's ability to satisfy users
    - System-oriented measures allow system developers to obtain an overall picture of performance
  - The metrics have been implemented in a Java package, EvalJ, which has been made available to all INEX participants as open source code on sourceforge.net

INEX 2005:
Overall Results

- **Stimulation of research activity** in new, previously unexplored areas, such as interactive aspects of XML information access and multimedia XML access

- Study and implementation of **evaluation methodologies** for diverse types of XML access systems

- Novel evaluation methodology where **user aspects** can be formalised in the evaluation measures (XCG metrics)

- Development of **test collections** and evaluation methodologies for XML access system evaluation

- Building of a **strong, multidisciplinary research community**

Plans for INEX 2006

INEX 2005 Tracks

- XML retrieval (ad-hoc)
- Interactive track
- Heterogeneous track
- Relevance feedback track
- Natural language query processing track
- Multimedia XML retrieval track
- Document mining track

- **New tracks**
  - XML entity search track (under discussion)

Details of new tracks and new tasks and evaluation methodologies under discussion

Further development of the infrastraucture for results analysis and testing of methodology