Overview of INEX 2005

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

[1] Information Systems, University of Duisburg-Essen, Duisburg, Germany
{malik,fuhr}@is.informatik.uni-duisburg.de
[2] Department of Computer Science, Queen Mary, University of London, London, UK
{gabs,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 2005.

1   Introduction

The INitiative for the Evaluation of XML retrieval (INEX) has, since 2002, 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. As a result of a collaborative effort, during the course of 2005, the INEX test collection has been further extended with an additional 4 712 new scientific articles from publications of the IEEE Computer Society, 87 new topics, and relevance judgements for 63 of these topics. Using the constructed test collection and the developed set of measures, the retrieval effectiveness of the participating organisations were evaluated and their results compared.

2005 has brought with it a lot of changes and new aspects to the evaluation. Several new tracks and tasks, a new relevance assessment procedure and new evaluation measures [2] were introduced. This paper presents an overview of these aspects and describes the work carried out as part of INEX 2005.

Section 2 gives a brief summary of this year's participants. Section 3 provides an overview of the expanded test collection. Section 4 outlines the retrieval tasks in the main ad-hoc track. Section 5 briefly reports on the submission runs for the ad hoc retrieval tasks. Section 6 describes the relevance assessment phase. The different measures used to evaluate retrieval performance are described in a separate paper [2]. Section 7 provides a short description of the tracks of INEX 2005. The paper wraps up with conclusions and outlook to INEX 2006.

2   Participating organisations

In response to the call for participation, issued in March 2005, 35 old and 12 new organizations registered. However throughout the year a number of groups dropped out due to resource requirements, while 11 further groups joined the initiative. The final 41 active groups along with their participation details are summarised in Table 1.

Table 1. List of active INEX 2005 participants

Organisations	Submitted topics	Assessed topics	Submitted runs
Max-Planck-Institut für Informatik	1	4	10
Royal School of LIS	5	2	0
University of California, Berkeley	2	2	20
University of Granada	4	2	0
University of Amsterdam	4	2	18
University of Otago	6	2	0
Queen Mary, University of London	0	2	0
University of Toronto	5	2	0
Utrecht University	6	2	16
City University London and Microsoft Research Cambridge	5	2	6
University of Kaiserslautern	3	2	14
IRIT	5	2	26
RMIT University	6	2	26
École Nationale Supérieure des Mines de Saint-Etienne	6	2	0
Queensland University of Technology	4	2	28
Universtity of Klagenfurt (ISYS)	0	2	3
University of Tampere	4	2	17
Carnegie Mellon University	3	2	4
University of Illinois at Urbana-Champaign	7	2	0
IBM Haifa Research Lab	6	2	26
University of Minnesota Duluth	8	2	24
Universidade Estadual de Montes Claros	4	2	8
The Hebrew University of Jeru	6	2	14
UCLA	6	2	2
University of Udine	0	2	0
VALORIA Lab, University of South-Brittany	0	2	0
Nagoya University	6	2	0
Laboratoire d'Informatique de Paris 6 (LIP6)	4	2	17
University of Waterloo	2	2	7
Kyungpook National University	0	2	9
University of Helsinki	0	2	7
Cirquid Project (CWI and University of Twente)	6	2	16
Universität Duisburg-Essen	1	1	0
Oslo University College	2	2	5
Universidad de Chile	0	1	0

Organizations participating only in the XML document mining track

INRIA

Charles de Gaulle University

University of Wolongong

Organization participating only in the interactive track

Rutgers University

3   The test collection

Test collections, as traditionally used in the information retrieval (IR), 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 IR consists of the same three parts, each component is rather different from its traditional IR counterpart.

Table 2. New additions to the IEEE collection in INEX 2005

id	Publication title	Year	Size (Mb)	No. of articles
an	IEEE Annals of the History of Computing	2002-2004	5.1	118
cg	IEEE Computer Graphics and Applications	2002-2004	7.6	220
co	Computer	2002-2004	14.8	664
cs	Computing in Science & Engineering	2002-2004	6.4	219
ds	IEEE Distributed Systems Online	2004	0.6	39
dt	IEEE Design & Test of Computers	2002-2004	6.1	263
ex	IEEE Intelligent Systems	2002-2004	8.2	240
ic	IEEE Internet Computing	2002-2004	7.0	264
it	IT Professional	2002-2004	3.4	142
mi	IEEE Micro	2002-2004	5.2	195
mu	IEEE Multimedia	2002-2004	4.6	161
pc	IEEE Pervasive Computing	2002-2004	5.1	160
so	IEEE Software	2002-2004	7.6	341
sp	IEEE Security & Privacy	2003-2004	4.4	179
tb	IEEE Transactions On Computational Biology & Bioinformatics	2004	0.8	12
tc	IEEE Transactions on Computers	2002-2004	27.5	319
td	IEEE Transactions on Parallel & Distributed Systems	2002-2004	23.2	235
tg	IEEE Transactions on Visualization and Computer Graphics	2002-2004	9.2	109
tk	IEEE Transactions on Knowledge and Data Engineering	2002-2004	26.9	255
tm	IEEE Transactions On Mobile Computing	2002-2004	6.5	79
tp	IEEE Transactions on Pattern Analysis and Machine Intelligence	2002-2004	28.9	350
tq	IEEE Transactions On Dependable and Secure Computing	2002-2004	1.1	12
ts	IEEE Transactions of Software Engineering	2002-2004	18.4	192
	Total new XML content added in INEX 2005 (incl. volume files):		228.6	4 768

In IR test collections, documents are considered 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, on the other hand, 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), represents a retrievable unit. In addition, with the use of XML query languages, users of an XML IR 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. These three components of the INEX test collection are described in the next sections.

3.1 Documents

This year the collection of documents that forms the INEX ad-hoc test collection has been extended with further publications donated by the IEEE Computer Society. The additional new resources are summarised in table 2. A total of 4 712 new articles (excluding the 56 new volume.xml files) from the period of 2002-2004 have been added to the previous collection of 12 107 articles, giving a total of 16 819 articles. This meant that the INEX ad-hoc test collection grew by 228Mb in size to a total of 764Mb.

3.2 Topics

As in previous years, INEX 2005 distinguished two basic types of topics: Content-Only (CO) and Content-And-Structure (CAS) topics. These topic types reflect two types of users with varying levels of knowledge about the structure of the searched collection. The first type simulates ignorant users, who either do not have any knowledge of the document structure or who choose not to use such knowledge. This profile is likely to fit most users searching XML digital libraries. The latter type of user aims to make use of any insight about the document structure that they may possess. They may then use this knowledge as a precision enhancing device in trying to make the information need more concrete. This user type is more likely to fit librarians.

Building on these basic types, INEX 2005, defined and investigated various extensions and interpretations of topic types.

Content-Only + Structure (CO+S). In an effort to investigate the usefulness of structural hints, the Content-Only (CO) topics, as used in previous years of INEX, were extended into so-called Content-Only + Structure (CO+S) topics. The aim was that the use of these topics enabled the comparison of a system's performance across two retrieval scenarios (on the same topic): when structural hints are taken into account (+S) and when these hints are ignored (CO).

As in previous years, queries with content-only conditions (CO queries) were defined as requests that ignore the document structure and contain only content related conditions, e.g. only specify what an element should be about without specifying what that component is. The topic format of CO queries includes a topic title, description and narrative.

The extended CO+S topics in INEX 2005 included an optional field called CAS title, which is a representation of the same information need but including additional knowledge in the form of structural hints (see the discussion on Topic format in this section).

```
<inex_topic topic_id="231" query_type="CO+S" ct_no="98" >
 <title>markov chains in graph related algorithms</title>
 <castitle>
    //article//sec[about(.,+"markov chains" +algorithm +graphs)]
 </castitle>
 <description>Retrieve information about the use of markov chains
    in graph theory and in graphs-related algorithms.
 </description>
 <narrative>I have just finished my MSc. in mathematics, in the field
    of stochastic processes. My research was in a subject related to
    Markov chains. My aim is to find possible implementations of my
    knowledge in current research. I'm mainly interested in
    applications in graph theory, that is, algorithms related to
    graphs that use the theory of markov chains. I'm interested in
    at least a short specification of the nature of implementation
    (e.g. what is the exact theory used, and to which purpose),
    hence the relevant elements should be sections, paragraphs or
    even abstracts of documents, but in any case, should be part of
    the content of the document (as opposed to, say, vt, or bib).
 </narrative>
</inex_topic>
```

Fig. 1. A CO+S topic from the INEX 2005 test collection

Content-And-Structure (CAS). The aim of the Content-And-Structure (CAS) topics this year was to support investigations on the different possible interpretations of structural constraint within a query, i.e. strict or vague, and the effect of this interpretation on retrieval effectiveness.

The actual definition of CAS topics have not changed from previous years: 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). More precisely, a CAS query contains two kinds of structural constraints: where to look (i.e. the support elements), and what to return (i.e. the target elements).

What was new in INEX 2005, was the explicit nature in which structural constraints were to be interpreted by a search system. Each structural constraint could be considered as a strict (must be matched exactly) or vague (simply as hints) criterion. The former closer reflects the database-oriented view, where only records that exactly match the specified structure should be returned to the user. The latter is closer to the IR view, where users' information need is assumed to be inherently uncertain. Four combinations of vague and strict interpretations of the structural constraints are then possible, depending on how the target elements and/or the containment conditions are treated:

– VVCAS: where the structural constraints in both the target elements and the support elements are interpreted as vague.
– SVCAS: where the structural constraints in the target elements are interpreted as strict and the structural constraints in the support elements are interpreted as vague.

- VSCAS: where the structural constraints in the target elements are interpreted as vague and the structural constraints in the support elements are interpreted as strict.
- SSCAS: where the structural constraints in both the target elements and the support elements are interpreted as strict.

```
<inex_topic topic_id="269"  query_type="CAS" ct_no="117" >
 <title> </title>
 <castitle>
    //article[about(.,interconnected networks)]//p[about(.,
    Crossbar networks)]
 </castitle>
 <description>We are looking for paragraphs that talk about
    Crossbar networks from articles that talk about interconnected
    networks.
 </description>
 <narrative>With networking between processors gaining significance,
    interconnected networks has become an important concept.
    Crossbar network is one of the interconnected networks. We are
    looking for information on what crossbar networks exactly are,
    how they operate and why they are used to connect processors.
    Any article discussing interconnected networks in the context
    of crossbar networks is considered to be relevant. Articles
    talking about interconnected networks such as Omega networks
    are not considered to be relevant. This information would be
    used to prepare a presentation for a lecture on the topic, and
    hence information on crossbar networks makes an element relevant.
 </narrative>
</inex_topic>
```

Fig. 2. A CAS topic from the INEX 2005 test collection

Topic format. Both CO+S and CAS topics are made up of several parts, each representing the same information need, but for different purposes.

- **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 an optional $+$ or $-$ prefix, where $+$ is used to emphasize an important concept, and $-$ is used to denote an unwanted concept.
- **CAS Title (castitle):** A short explanation of the information need, specifying any structural requirements. The CAS title is optional in CO+S topics, but mandatory in CAS topics. Similarly to topic title, a word in the CAS title can have a $+$ or $-$ prefix. CAS titles are expressed in the query language of NEXI [5].
- **Parent:** Only used for CAS topics. Each CAS topic containing more than one `about` function was submitted with a set of sub-topics describing the information need of each single `about` clause. In order to match the sub-topics with the topic the parent had to be identified in the sub-topic.

- **Description:** a one or two sentence natural language definition of the information need.
- **Narrative:** a detailed explanation of the information need and a 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 7.1).

The title and the description had to be interchangeable. This was a requirement of the natural language processing track (see Section 7.4). The DTD of the topics is shown in Figure 3.

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

Fig. 3. Topic DTD

Attributes of a topic are: topic_id (which in INEX 2005 ranges from 202 to 288), query_type (with possible values of "CAS" or "CO+S") and ct_no, which refers to the candidate topic number (ranging from 1 to 145[3]). Examples of both types of topics are given in Figures 1 and 2.

Topic creation. Topics were created by the participating groups. Each group was asked to submit up to 6 candidate topics (3 CO+S and 3 CAS). A detailed guideline was provided to the participants for the topic creation [7].

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 regard to system capabilities or collection peculiarities to avoid artificial or collection

---

[3] Note that, due to the withdrawal of some topics, this is not a continuous range.

Table 3. Statistics on CAS and CO+S topics on the INEX 2005 test collection

	CAS	CO+S
Number of topics	40	47
Average length of title (in words)	-	3.8
Boolean operators (and/or) in title	-	44
Prefix operators (+/-) in title	-	7
Phrases in title	-	20
Boolean operators (and/or) in castitle	5	13
Prefix operators (+/-) in castitle	2	9
Phrases in castitle	52	17
Average length of topic description (in words)	13	17
Average length of narrative (in words)	73	91

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 made available to participants to help with this task. Participants were asked to judge the top 25 retrieved results and record for each found relevant document/component its file name and its XPath. Those topics having at least 2 relevant documents/components but less than 20 documents/components were to 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.

After the completion of the first three stages, topics were submitted to INEX. A total of 139 candidate topics were received, of which 87 topics (40 CO+S and 47 CAS) were selected to form the final set of topics added to the test collection. Topic selection was based on a combination of criteria such as 1) balancing the number of topics across all participants, 2) eliminating topics that were considered too ambiguous or too difficult to judge, 3) uniqueness of topics, and 4) considering their suitability to the different tracks. Table 3 shows some statistics on the final set of INEX 2005 topics.

4 Retrieval tasks

The main retrieval task at INEX 2005 was defined as the ad-hoc retrieval of XML documents [8]. In information retrieval literature, ad-hoc retrieval is described as a simulation of how a library might be used. It involves the searching of a static set of documents using a new set of topics. While the principle is the same, the difference for INEX is that the library consists of XML documents, the queries may contain both content and structural conditions and, in response to a query, arbitrary XML elements may be retrieved from the library.

Unlike previous years, INEX 2005 distinguished several retrieval strategies, each based on different assumptions regarding a search system's output. These strategies have been explicitly investigated within the ad-hoc sub-tasks that build on the use of CO and CO+S queries. For tasks based on the use of CAS queries, systems' were assumed to follow the Thorough retrieval strategy only.

4.1 Ad-hoc sub-tasks based on CO queries

CO.Focussed: This strategy was intended for approaches concerned with the focussed retrieval of XML elements, i.e. aiming at targeting the appropriate level of granularity of relevant content that should be returned to the user for a given topic. An explicit assumption here was that a retrieval run should not contain any overlapping elements. The aim was for systems to find the most exhaustive and specific element on a path within a given document containing relevant information and return to the user only this most appropriate unit of retrieval. In the case where an XML retrieval system has estimated a parent and one if its child elements to be equally exhaustive and specific for a given topic, the parent element were to be returned. In addition, when a parent has been estimated as more exhaustive than one of its child elements, but the child element has been estimated as more specific than its parent, then the child element was to be selected. In this way, preference for specificity over exhaustivity was given.

CO.Thorough: This strategy was intended for XML retrieval approaches that do not deal with the problem of overlap when generating their output list for the evaluation, but consider this an interface and results presentation issue, which is to be resolved at a later stage, outside the scope of the evaluation. The aim here was for systems to find all relevant elements within the collection. Due to the nature of relevance in XML retrieval (e.g. if a child element is relevant, so will be its parent, although to a greater or lesser extent), an XML retrieval system that has estimated an element to be relevant may decide to return all its ancestor elements. This means that runs for this task may contain a large number of overlapping elements. It is however a challenge to rank these elements appropriately, as systems that rank highly exhaustive and specific elements before less exhaustive and specific ones were to obtain better effectiveness.

CO.FetchBrowse: This strategy was intended for XML retrieval approaches that are based on a mixture of document retrieval and element retrieval strategies. The aim of the fetch and browse retrieval strategy was to first identify relevant articles (the fetching phase), and then to identify the most exhaustive and specific elements within the fetched articles (the browsing phase). In the fetching phase, articles had to be ranked according to how exhaustive and specific they were (i.e. the most exhaustive and specific articles were to be ranked first). In the browsing phase, ranking had to be done according to how exhaustive and specific the relevant elements in the article were, compared to other elements in the same article.

4.2 Ad-hoc sub-tasks based on CO+S queries

Upon discovering that a CO query returned many irrelevant hits, a user may decide to add structural hints in order to improve precision. These structural hints were expressed in the `<castitle>` part of the CO+S topics, which was then used as the query for the CO+S sub-tasks. The aim of the CO+S sub-tasks was to specifically investigate the usefulness of the structural hints. As for the CO sub-tasks, three retrieval strategies were defined: COS.Focussed, COS.Thorough and COS.FetchBrowse.

4.3 Ad-hoc sub-tasks based on CAS queries

As described in section 3.2, different interpretations of CAS topics on the basis of target and support elements were possible, resulting in the sub-tasks of VVCAS, SVCAS, VSCAS and SSCAS. In these sub-tasks, the aim was to retrieve the most exhaustive and specific elements with respect to the topic of request, where overlap among retrieval results was allowed (Thorough strategy). An analysis of the outcome of the CAS sub-tasks can be found in a separate paper [4].

5 Submissions

During the retrieval runs, participating organisations evaluated the 87 INEX 2005 topics (40 CO+S and 47 CAS queries) against the IEEE Computer Society document collection and produced a list (or set) of document components (XML elements) as their retrieval results for each topic. The top 1500 components in a topic's retrieval results were then submitted to INEX. Table 4 summarises the submissions to the different ad-hoc tasks. For each topic, around 500 articles along with their components were pooled from all the submissions in round robin way for relevance assessment. Table 5 shows the pooling effect on the CAS and CO+S topics.

Table 4. Number of runs submitted to the different ad-hoc tasks

CO.Focussed	44	COS.Focussed	27
CO.Thorough	55	COS.Thorough	33
CO.FetchBrowse	42	COS.FetchBrowse	25
VVCAS	28	SSCAS	25
VSCAS	23	SVCAS	23

Table 5. Pooling effect for CAS and CO+S topics

	CAS topics	CO+S topics
Number of articles submitted	176 735	236 060
Number of articles in pools	23 250	20 135
Number of components submitted	812 207	1 337 214
Number of components in pools	92 905	80 019

6 Assessments

6.1 Relevance dimensions and scales

Relevance assessments were given according to two relevance dimensions [9]:

- **Exhaustivity (e)**, which describes the extent to which the document component discusses the topic of request.
- **Specificity (s)**, which describes the extent to which the document component focuses on the topic of request.

While the above definition of the relevance dimensions has remained unchanged since 2003, the scale that these dimensions were measured on has been revised in INEX 2005. The scale for exhaustivity was changed to $3 + 1$ levels: highly exhaustive ($e = 2$), somewhat exhaustive ($e = 1$), not exhaustive ($e = 0$) and "too small" ($e = ?$). The latter category of "too small" was introduced with the aim to allow assessors to label document components, which although contained relevant information were too small to sensibly reason about their level of exhaustivity[4]. Specificity was measured automatically on a continuous scale with values in [0, 1], where s = 1 represents a fully specific component (i.e. contains only relevant information). Values of specificity were derived on the basis of what ratio of a document component has been highlighted by the assessor (see section 6.2).

We denote the relevance degree of an assessed component, given by the combined values of exhaustivity and specificity, as (e, s), where e $\in \{?, 0, 1, 2\}$ and s $\in$ [0, 1]. For example, (2, 0.72) denotes a highly exhaustive component, 72% of which is relevant content.

6.2 Relevance assessments procedure

A relevance assessment guideline explaining the relevance dimensions and how and what to assess was distributed to the participants [9]. This guide also contained the manual to the online assessment tool, developed by Benjamin Piwowarski. The tool is referred to as X-RAI (XML Retrieval Assessment Interface - see Figures 4 and 5).

In order to reduce assessment effort, a highlighting procedure was used in INEX 2005 leading to the following process for assessment:

- In the first pass, assessors were asked to highlight text fragments that contained only relevant information - see Figure 5.
- In the second pass, assessors judged the exhaustivity level of any elements that had highlighted parts.

As a result of this process, any elements that have been fully highlighted were automatically labeled as fully specific ($s = 1$). The main advantage of this highlighting approach was that assessors now only had to judge the exhaustivity level of the elements that have highlighted parts (in the second phase). The specificity of any other (partially highlighted) elements was calculated automatically as a function of the contained relevant and irrelevant content, and more specifically, as the ratio of relevant

---

[4] The notion of "too small" has originally been employed in INEX 2002, there as a degree of coverage. It was removed from subsequent INEX evaluations as it was showed that assessors often labeled descendant components of target elements in CAS queries as "too small". Its re-introduction into the evaluation, but this time, more appropriately, as a degree of exhaustivity, was deemed necessary in order to free assessors from the burden of having to assess very small text fragments whose level of exhaustivity could not be sensibly decided.

content to all content, measured in number of words or characters. All non-highlighted elements were automatically assumed as not exhaustive ($e = 0$).

6.3 CAS assessments

This year there were four sets of CAS judgments, one for each of the four CAS interpretations - each derived from the same initial set of judgments. These are described in [4].
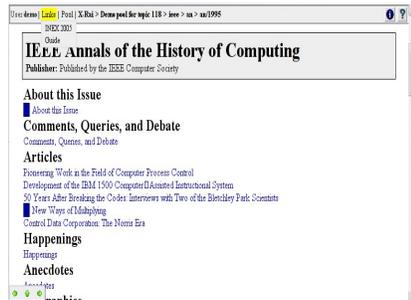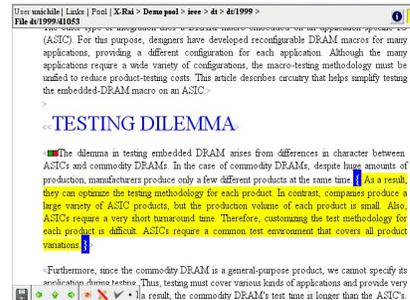

Fig. 4. X-RAI in assessment mode

Fig. 5. X-RAI Article view

7 INEX 2005 Tracks

In addition to the main ad hoc track, six research tracks were included in INEX 2005, each studying different aspects of XML information access: Interactive, Relevance Feedback, Heterogeneous, Natural Language Processing, and two new tracks for 2005, Multimedia and Document Mining tracks.

7.1 Interactive Track

In its second year, the Interactive Track (iTrack) focused on addressing some fundamental issues of interactive XML retrieval. In addition, the track also expanded by including two additional tasks and by attracting more participating groups. A total of 11 research groups and 108 test persons participated in the three different tasks that were included in the track. Details of the track can be found in [3].

7.2 Relevance Feedback track

The Relevance Feedback track investigated approaches for queries that also include structural hints (rather than content-only queries as in 2004). To limit the number of submissions, a subset of ad-hoc tasks were chosen for participants to test their

relevance feedback algorithms. These included CO.Thorough, CO+S.Thorough and VVCAS tasks. The reported evaluation scores for each relevance feedback submission measured the relative and absolute improvement of the relevance feedback run over the original base run. Five groups submitted 15 runs for CO.Thorough task, 9 runs for COS.Thorough task and 3 runs for VVCAS task.

7.3 Heterogeneous track

The Heterogeneous track expanded by studying new collections with different DTDs and their effect on XML IR system effectiveness. 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): 3 465 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): 2 024 XML items.
– ZDNet (Articles and Comments) provided by ZDNet.com to the INEX evaluation: 96 351 items (4 734 Articles and 91 617 comments on those articles). This sub-collection was added in 2005.

7.4 Natural Language Processing track

The Natural Language Processing track (NLPX) focused on whether it is possible to express topics in natural language, which is then to be used as basis for retrieval. For this year, two tasks were defined NLQ2NEXI and NLQ. NLQ2NEXI required the translation of a natural language query, provided in the element of a topic, into a formal INEX <title> query element. The NLQ task had no restrictions on the use of any NLP techniques to interpret the queries as they appeared in the <description> element of a topic. The objective was not only to compare between different NLP based systems, but to also compare the results obtained with natural language queries with the results obtained with NEXI queries by any other system in the ad hoc track. During the topic creation stage, it was ensured that the description component of the topics were equivalent in meaning to their corresponding NEXI title, so it was possible to re-use the same topics, relevance assessments and evaluation procedures as in the ad hoc track. The topic descriptions were used as input to natural language processing tools, which processed them into representations suitable for XML search engines. Three groups submitted 12 runs for CO.Focussed task, 5 runs for COS.Thorough task, 5 runs for COS.FetchBrowse and 8 runs for CAS tasks. The results showed that NLQ2NEXI task performed better than the NLQ task.

7.5   Multimedia track

The main objective of the Multimedia track was to provide a pilot evaluation platform and 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. Full details of the track can be found in [6].

7.6   Document Mining track

The aim of the Document Mining track, run in collaboration with the PASCAL network of Excellence[5], was 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. Two new collections were developed: The WIPO corpus which is composed of 75 250 XML documents, and the MovieDB corpus (based on the Internet Movie Database) which consists of 9 463 XML documents.

8   Conclusion and outlook

INEX 2005 has shown that XML retrieval is a challenging field. In addition to learning more about XML retrieval approaches, INEX 2005 has introduced several new aspects and made several changes to the evaluation methodology:

– The document collection was extended to include now 16 819 articles of the IEEE Computer Society's publications, increasing the size of the collection to a total of 764Mb (containing over 10 million XML elements). A number of new document collections were also added and used in the various tracks. For example, the multimedia track conducted experiments using the Lonely Planet WorldGuide XML collection.
– A new assessment procedure was introduced with the aim to reduce assessment effort (both with respect to cognitive load and time required).
– A range of ad-hoc retrieval tasks were investigated with the aim to address specific research questions, e.g. the impact of structure as precision enhancing device or the interpretation of structural query constraints.
– In addition to the ad-hoc retrieval tasks, several retrieval strategies were studied, each based on different assumptions regarding what users would want to obtain as the outcome of a search.
– A new set of evaluation measures were employed with the aim to address problematic issues encountered with precision-recall like metrics.
– INEX 2005 run a total of 7 tracks, each studying different aspects of XML information access: Ad-hoc, Interactive, Relevance Feedback, Heterogeneous, Natural Language Processing, and two new tracks for 2005, Multimedia and Document Mining tracks.

---

[5] http://www.pascal-network.org/

INEX 2006 will start in March of this year. In addition to the current tracks, INEX 2006 will have two new tracks: User case studies and XML entity ranking tracks. In addition, a new test collection will be used, based on the Wikipedia project. Statistical analysis of the various measures employed are also currently ongoing; results of this will provide input for selecting which of these measures to use in 2006.

References

1. Norbert Fuhr, Norbert Gövert, and Kai Großjohann. HyREX: Hyper-media retrieval engine for XML. In *Proceedings of the 25th Annual International Conference on Research and Development in Information Retrieval*, page 449, 2002. Demonstration.
2. Gabriella Kazai and Mounia Lalmas. INEX 2005 evaluation metrics. In *Advances in XML Information Retrieval and Evaluation: Fourth Workshop of the INitiative for the Evaluation of XML Retrieval (INEX 2005), Lecture Notes in Computer Science, Vol 3977, Springer-Verlag*, 2006.
3. Birger Larsen, Saadia Malik, and Anastasios Tombros. The interactive track at INEX 2005. In *Advances in XML Information Retrieval and Evaluation: Fourth Workshop of the INitiative for the Evaluation of XML Retrieval (INEX 2005), Lecture Notes in Computer Science, Vol 3977, Springer-Verlag*, 2006.
4. Andrew Trotman and Mounia Lalmas. The Interpretation of CAS. In *Advances in XML Information Retrieval and Evaluation: Fourth Workshop of the INitiative for the Evaluation of XML Retrieval (INEX 2005), Lecture Notes in Computer Science, Vol 3977, Springer-Verlag*, 2006.
5. Andrew Trotman and Börkur Sigurbjörnsson. Narrowed extended XPATH I (NEXI). In *Advances in XML Information Retrieval, Third Workshop of the Initiative for the Evaluation of XML Retrieval (INEX 2004), Dagstuhl, Germany, December 6-8, 2004, Revised Selected Papers, Lecture Notes in Computer Science, Vol 3493, Springer-Verlag*, page 16–40, 2005.
6. Roelof van Zwol, Gabriella Kazai, and Mounia Lalmas. INEX 2005 multimedia track. In *Advances in XML Information Retrieval and Evaluation: Fourth Workshop of the INitiative for the Evaluation of XML Retrieval (INEX 2005), Lecture Notes in Computer Science, Vol 3977, Springer-Verlag*, 2006.
7. Börkur Sigurbjörnsson, Andrew Trotman, Shlomo Geva, Mounia Lalmas, Birger Larsen, and Saadia Malik INEX 2005 Guidelines for Topic Development. In *INEX 2005 Workshop Pre-Proceedings, Dagstuhl, Germany, November 28–30, 2005*, page 375–384, 2005.
8. Mounia Lalmas INEX 2005 Retrieval Task and Result Submission Specification. In INEX 2005 Workshop Pre-Proceedings, Dagstuhl, Germany, November 28–30, 2005, page 385–390, 2005.
9. Mounia Lalmas and Benjamin Piwowarski INEX 2005 Relevance Assessment Guide. In INEX 2005 Workshop Pre-Proceedings, Dagstuhl, Germany, November 28–30, 2005, page 391–400, 2005.