

Overview of INEX 2004

Saadia Malik¹, Mounia Lalmas², and Norbert Fuhr³

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 Retrieval¹ (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/>

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 no of runs	
	topics	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 Bordon	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>

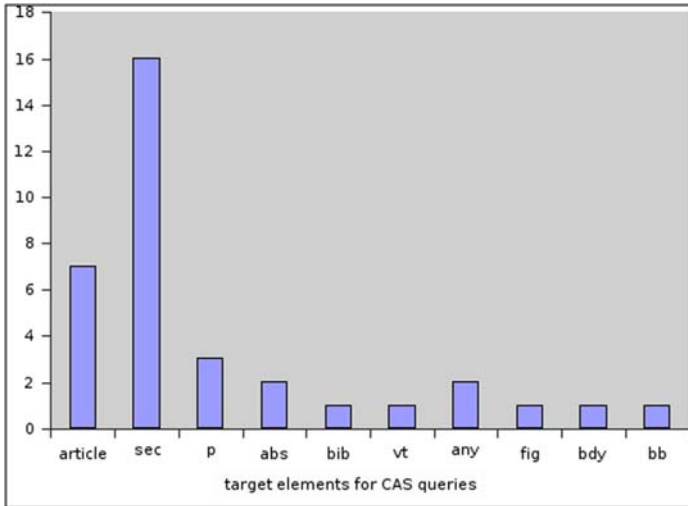
```

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.

Table 2. Statistics on CAS and CO topics on the INEX test collection

	CAS	CO
no of topics	34	39
avg no of words in title	11	5
no of times +/- used	12	52
avg no of words in topic description	20	23
avg no of words in keywords component	6	7
avg no of words in narrative component	67	98

**Fig. 4.** Target elements in CAS topics

After the completion of the first three stages, topics were submitted to INEX. A total of 198 candidate topics were received, of which 73 topics (39 CO and 34 CAS) were selected. The topic selection was based on the basis of 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 and 3) uniqueness of topics, and 4) considering their suitability to the different tracks. Table 2 shows some statistics on the INEX 2004 topics². Figure 4 shows the distribution of target elements in the CAS topics.

4 The Retrieval Tasks

The retrieval task to be performed by the participating groups at INEX 2004 was defined as the ad hoc retrieval of XML documents. In information retrieval (IR) literature, ad

² A word in the title component of the topic can have either the prefix + or -, where + is used to emphasize an important concept, and - is used to denote an unwanted concept.

hoc retrieval is described as a simulation of how a library might be used, and 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. Within the ad hoc retrieval task, INEX 2004 defined the following two sub-tasks: CO and VCAS.

The **CO task** stands for content-oriented XML retrieval using CO queries. The elements to retrieve are components that are most specific and most exhaustive with respect to the topic of request. Most specific here means that the component is highly focused on the topic, while exhaustive reflects that the topic is exhaustively discussed within the component.

The **VCAS task** stands for content-oriented XML retrieval based on CAS queries, where the structural constraints of a query can be treated as vague conditions. The idea behind the VCAS sub-task was to allow the evaluation of XML retrieval systems that aim to implement approaches, where not only the content conditions within a user query are treated with uncertainty but also the expressed structural conditions. The structural conditions were to be considered hints as to where to look.

5 Submissions

Participants processed the final set of topics with their retrieval systems and produced ranked lists of 1500 result elements in a specified format. Participants could submit up to 3 runs per sub-task, CO and VCAS. In total 121 runs were submitted by 26 participating organisations. Out of the 121 submissions, 70 contained results for the CO topics, and 51 contained results for the CAS topics. For each topic, around 500 articles along with their components were pooled from all the submissions in round robin way for assessment. Table 3 shows the pooling effect on the CAS and CO topics.

Table 3. Pooling effect for CAS and CO topics

	CAS topics	CO topics
no of documents submitted	160264	200988
no of documents in pools	17092	19640
no of components submitted	677022	858724
no of components in pools	45244	45280

6 Assessments

The assessment pools were assigned then to participants; either to the original authors of the topic when this was possible, or on a voluntary basis, to groups with expertise in the topic's subject area. Each group was responsible for about two topics. In order to obtain some duplicate assessments, 12 topics were assigned to two participants, thus resulting in two sets of relevance assessments, referred to as Ass. I and Ass. II here.

Since 2003, relevance in INEX is defined according to the following two dimensions:

- **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.

Table 4. Assessments at article and component levels with Assessment (Ass.) set I and Assessment set II

e+s	VCAS Ass. I		VCAS Ass. II		CO Ass. I		CO Ass. II	
	article	non-article	article	non-article	article	non-article	article	non-article
e3s3	0.87%	0.85%	0.91%	0.82%	0.98%	0.61%	1.10%	0.63%
e3s2	0.74%	0.16%	0.95%	0.16%	0.89%	0.49%	0.98%	0.51%
e3s1	7.88%	0.24%	8.29%	0.25%	1.61%	0.12%	1.64%	0.13%
e2s3	0.26%	0.34%	0.39%	0.38%	0.41%	0.66%	0.41%	0.72%
e2s2	0.73%	0.26%	1.02%	0.30%	0.49%	0.86%	0.63%	0.92%
e2s1	2.76%	0.18%	3.04%	0.19%	2.07%	0.30%	2.23%	0.31%
e1s3	0.19%	1.62%	0.25%	1.97%	0.47%	0.92%	0.45%	1.23%
e1s2	0.84%	0.46%	1.11%	0.61%	0.34%	0.57%	0.35%	0.67%
e1s1	8.36%	1.46%	9.42%	1.48%	5.20%	1.28%	6.10%	1.47%
e0s0	77.36%	94.42%	74.63%	93.84%	87.55%	94.20%	86.10%	93.41%
All	100.00%	100.00%	100.00%	100.00%	100.00%	100.00%	100.00%	100.00%

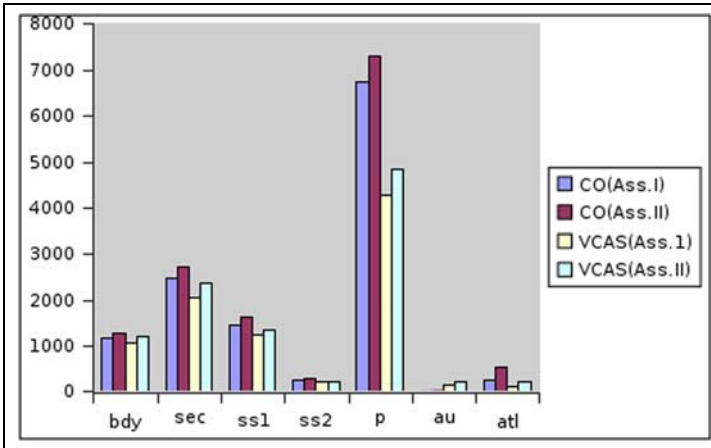


Fig. 5. Distribution of relevant elements

Exhaustivity was measured on the following 4-point scale: Not exhaustive (e0): the document component does not discuss the topic of request at all; Marginally exhaustive (e1): the document component discusses only few aspects of the topic of request; Fairly exhaustive (e2): the document component discusses many aspects of the topic of request; and Highly exhaustive (e3): the document component discusses most or all aspects of the topic of request.

Specificity was assessed on the following 4-point scale: Not specific (s0): the topic of request is not a theme of the document component; Marginally specific (s1): the topic of request is a minor theme of the document component (i.e. the component focuses on other, non-relevant topic(s), but contains some relevant information); Fairly specific (s2): the topic of request is a major theme of the document component (i.e. the component contains mostly relevant content and only some irrelevant content); and Highly specific (s3): the topic of request is the only theme of the document component.

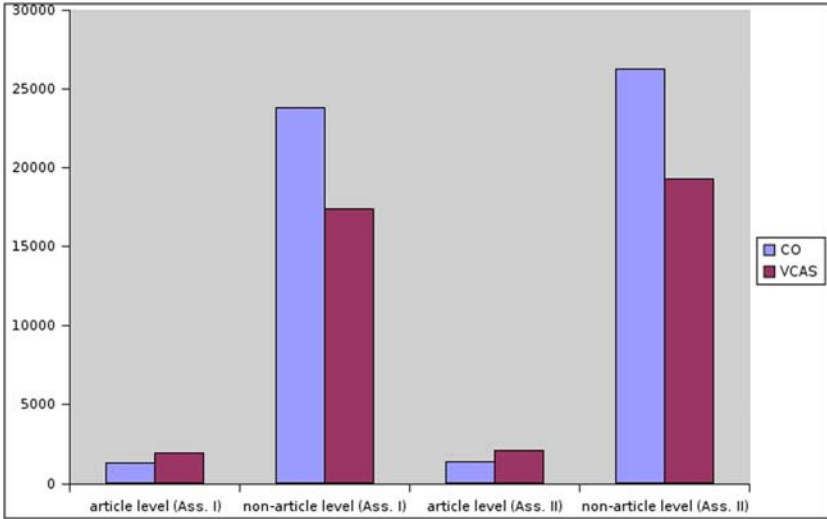


Fig. 6. Distribution of relevant article and non-article elements ($e > 0$ and $s > 0$)

Although the two dimensions are largely independent of each other, a not-exhaustive ($e0$) component can only be not specific ($s0$) and vice versa. Other than this rule, a component may be assigned any other combination of exhaustivity and specificity, i.e. $e3s3$, $e3s2$, $e3s1$, $e2s3$, $e2s2$, $e2s1$, $e1s3$, $e1s2$, and $e1s1$. For example, a component assessed as $e1s1$ is one that contains only marginally exhaustive relevant information ($e1$) where this relevant content is only a minor theme of the component, i.e. most of the content is irrelevant to the topic of request ($s1$).

A relevance assessment guideline explaining the relevance dimensions and how and what to assess was distributed to the participants. This guide also contained the manual to the online assessment tool developed by LIP6 to perform the assessments of the XML documents/components. Features of the tool include user friendliness, implicit assessment rules whenever possible, keyword highlighting, consistency checking and completeness enforcement. Table 4 shows a statistics of the relevance assessments. Figures 5 and 6 show the distribution of relevance for (some of) the elements.

Initial investigations show that agreements on the relevance assessments is significantly lower than in other evaluation initiatives. However, this outcome is affected by two unique features of INEX, the multivalued relevance scale, and the nesting of result

elements in non-atomic documents. Further analysis is needed in order to quantify the effect of these factors, and their impact on the overall retrieval quality.

7 Evaluation Metrics

For evaluation the *inex_eval* (also referred to as *inex-2002*) metric is used with a number of quantization functions. This metric was developed during INEX 2002, and was adapted to deal with the two dimensions of relevance (i.e. exhaustivity and specificity) that were adopted in INEX 2003. *inex_eval* is based on the traditional recall and precision measures. To obtain recall/precision figures, the two dimensions need to be quantised onto a single relevance value. The following quantisation functions reflecting different user standpoints were used in INEX 2004:

- A strict quantisation (*strict*) evaluates whether a given retrieval approach is capable of retrieving highly exhaustive and highly specific document components (e3s3).
- A generalised quantisation (*general*) credits document components according to their degree of relevance. This quantisation favours exhaustivity over specificity.
- Specificity-oriented general (*sog*) credits document components according to their degree of relevance, but where specificity is favoured over exhaustivity.
- Exhaustivity-oriented (s_3e_{321} , s_3e_{32}) quantisations apply the strict quantisation with respect to the exhaustivity dimension and allow to consider different degrees of specificity.
- Specificity-oriented (s_3e_{321} , s_3e_{32}) quantisations apply the strict quantisation with respect to the specificity dimension and allow to consider different degrees of exhaustivity.

Based on the quantised relevance values, procedures that calculate recall/precision curves for standard document retrieval were directly applied to the results of the quantisation functions. The method of *precall* described by [5] was used to obtain the precision values at standard recall values. Further details are available in [4].

8 Summary of Evaluation Results

As mentioned in Section 5, out of the 121 submissions, 70 contained results for the CO task, and 51 contained results for the VCAS task. A summary of the results obtained with the evaluation metric (*inex_eval*) is given here³. The submissions have been ranked according to the average precision over all the quantisations (arithmetic mean). The top ten submissions for each task and with both assessment sets are listed in Table 5 and Table 6. Furthermore, an overlap indicator characterising a run by the percentage of overlapping items in the submission was also calculated.

³ All evaluation results have been compiled using the assessment package version 3.0 and evaluation package version 2004.003.

Table 5. Ranking of submissions for CO task; average precision over all quantisations and overlap indicator

rank organisation(run ID)	avg precision overlap(%)	
1. IBM Haifa Research Lab(CO-0.5-LAREFIENMENT)	0.1437	80.89
2. IBM Haifa Research Lab(CO-0.5)	0.1340	81.46
3. University of Waterloo(Waterloo-Baseline)	0.1267	76.32
4. University of Amsterdam(UAms-CO-T-FBack)	0.1174	81.85
5. University of Waterloo(Waterloo-Expanded)	0.1173	75.62
6. Queensland University of Technology(CO_PS_Stop50K_099_049)	0.1073	75.89
7. Queensland University of Technology(CO_PS_099_049)	0.1072	76.81
8. IBM Haifa Research Lab(CO-0.5-Clustering)	0.1043	81.10
9. University of Amsterdam(UAms-CO-T)	0.1030	71.96
10. LIP6(simple)	0.0921	64.29

a) Ranking based on assessment set I

rank organisation(run ID)	avg precision overlap(%)	
1. IBM Haifa Research Lab(CO-0.5-LAREFIENMENT)	0.1385	80.89
2. IBM Haifa Research Lab(CO-0.5)	0.1285	81.46
3. University of Amsterdam(UAms-CO-T-FBack)	0.1212	81.85
4. University of Waterloo(Waterloo-Baseline)	0.1195	76.32
5. University of Waterloo(Waterloo-Expanded)	0.1113	75.62
6. Queensland University of Technology(CO_PS_Stop50K_099_049)	0.1084	75.89
7. Queensland University of Technology(CO_PS_099_049)	0.1064	76.81
8. University of Amsterdam(UAms-CO-T)	0.1047	71.96
9. IBM Haifa Research Lab(CO-0.5-Clustering)	0.1016	81.10
10. LIP6(simple)	0.0967	64.29

b) Ranking based on assessment set II

9 INEX 2004 Tracks

INEX 2004 had four new tracks, which are briefly described in this Section.

9.1 Relevance Feedback Track

The aim of this track is to investigate relevance feedback in the context of XML retrieval. In standard full text search engines, relevance feedback (RF) has been translated into detecting a “bag of words” that are good (or bad) at retrieving relevant information. These terms are then added to (or removed from) the query and weighted according to their power in retrieving relevant information. With XML documents, a more sophisticated approach - one that can exploit the characteristics of XML - is necessary. The approach should ideally consider not only content but also the structural features of XML documents. The query reformulation process must therefore infer which content and structural elements are important for effectively retrieving relevant data.

For the first year of the RF track, the focus was on CO topics. The participant’s CO runs served as the baselines, upon which RF was performed - based on the relevance assessments - from the top-ranked 20 elements retrieved in the original CO run. There

Table 6. Ranking of submissions for VCAS task; average precision over all quantisations and overlap indicator

rank organisation(run ID)	avg precision overlap(%)	
1. Queensland University of Technology(VCAS_PS_stop50K_099_049)	0.1260	82.46
2. Queensland University of Technology(VCAS_PS_099_049)	0.1244	82.89
3. Queensland University of Technology(VCAS_PS_stop50K_049025)	0.1171	78.64
4. University of Amsterdam(UAms-CAS-T-FBack)	0.1065	77.76
5. IRIT(VTCAS2004TC35xp200sC-515PP1)	0.0784	76.33
6. Carnegie Mellon U.(Lemur_CAS_as_CO_NoStem_Mix02_Shrink01)	0.0759	74.00
7. Cirquid Project (CWI and U. of Twente)(LMM-VCAS-Relax-0.35)	0.0694	24.31
8. IBM Haifa Research Lab(CAS-0.5)	0.0685	38.76
9. Cirquid Project (CWI and U. of Twente)(LMM-VCAS-Strict-0.35)	0.0624	22.78
10. University of Amsterdam(UAms-CAS-T-XPath)	0.0619	18.78

a) Ranking based on assessment set I

rank organisation(run ID)	avg precision overlap(%)	
1. Queensland University of Technology(VCAS_PS_099_049)	0.1245	82.89
2. Queensland University of Technology(VCAS_PS_stop50K_049025)	0.1168	78.64
3. Queensland University of Technology(VCAS_PS_stop50K_099_049)	0.1159	82.46
4. University of Amsterdam(UAms-CAS-T-FBack)	0.1097	77.76
5. Carnegie Mellon U.(Lemur_CAS_as_CO_NoStem_Mix02_Shrink01)	0.0788	74.00
6. IBM Haifa Research Lab(CAS-0.9-LAREFIENMENT)	0.0781	43.52
7. IBM Haifa Research Lab(CAS-0.5)	0.0750	38.76
8. IRIT(VTCAS2004TC35xp200sC-515PP1)	0.0737	76.33
9. Cirquid Project (CWI and U. of Twente)(LMM-VCAS-Relax-0.35)	0.0700	24.31
10. University of Amsterdam(UAms-CAS-T-XPath)	0.0642	18.78

b) Ranking based on assessment set II

were no restrictions on the number of iterations of relevance feedback for a given query. Participants were allowed to submit at most three RF runs per initial ad hoc run. Three groups submitted 15 RF runs, where the best performing run was CO-0.5-ROCCHIO by IBM Haifa Research Lab.

One major issue was the evaluation methodology itself. Given that RF attempts to improve performance by using information in marked relevant documents (in the case of INEX the top 20 elements), it is usually the case that one of the main effects of the RF process is to push the known relevant documents to the top of the document ranking. This ranking effect, will artificially improve performance figures for the new document ranking simply by re-ranking the known relevant documents. What is not directly tested is how good the RF technique is at improving retrieval of unseen relevant documents - the feedback effect. There are two main alternatives to measure the effect of feedback on the unseen documents.

In the residual ranking methodology, the documents which are used in RF are removed from the collection before evaluation. This will include the relevant and some non-relevant documents (in our case the top 20 elements). After RF, the performance figures are calculated on the remaining (residual) collection. The problem with this ap-

proach is that the feedback results are not comparable with the original ranking. This is because the residual collection has fewer documents, and fewer relevant documents, than the original collection. A further difficulty is that, at each successive iteration of feedback, performance figures may be based on different numbers of queries because some queries may not have any relevant documents in the residual collection, and hence are removed. This implies that RF runs are not comparable across participants.

In the freezing methodology, the rank positions of the top n documents (our top 20 elements), the ones used to modify the query, are frozen. The remaining documents are re-ranked and performance figures are calculated over the whole ranking. As the only documents to change rank position are those below n (the ones used for RF) any change in performance happens as a result of the change of rank position of the unseen relevant documents. The freezing method was adopted in INEX 2004.

9.2 Heterogeneous Collection Track

The current INEX collection is based on a single DTD. In practical environments, such a restriction will hold in rare cases only. Instead, most XML collections will comprise documents from different sources, and thus with different DTDs. Also, there will be distributed systems (federations or peer-to-peer systems), where each node manages a different type of collection. The heterogeneous document collection track aims at investigating these issues, and in particular the following challenges:

- For content-only queries, most current approaches use the DTD for defining elements that would form reasonable answers. In heterogeneous collections, DTD-independent methods should be developed.
- For CAS queries, there is the problem of mapping structural conditions from one DTD onto other (possibly unknown) DTDs. Methods from federated databases could be applied here, where schema mappings between the different DTDs are defined manually. However, for a larger number of DTDs, automatic methods must be developed.

In the first year, the track was mainly explorative. The focus was on the construction of an appropriate test collection and the elaboration of the research issues and challenges. A collection comprised of several sub-collections was built. Twenty ad hoc topics (CO and CAS) were reused and four new topics were created. Three participants submitted 19 runs. The pooling, assessment, evaluation and browsing of the collection were shown to pose new challenges. Full details can be found in [6].

9.3 Natural Language Processing Track

The natural language processing track at INEX 2004 focused on whether it is possible to express topics in natural language, to be then used as basis for retrieval. During the topic creation stage, it was ensured that the description component of the topics were equivalent in meanings 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 descriptions were used as input to natural language processing tools, which would process them into representations suitable for XML search engines.

Four groups submitted 7 runs (4 CO and 3 VCAS). The best performing run for the CO task was CO-FRAGMENT-RANKING-NLP-0.5 by IBM Haifa Research Lab; and for the VCAS task, it was NLP_CAS_099_049_PS_100K by Queensland University of Technology.

9.4 Interactive Track

The main motivation for the track was twofold. First, to investigate the behaviour of users when interacting with components of XML documents, and secondly to investigate and develop approaches for XML retrieval which are effective in user-based environments.

In INEX 2004, only the first issue was addressed: to investigate the behaviour of searchers when presented with components of XML documents that have a high probability of being relevant (as estimated by an XML-based IR system). Presently, all metrics that are in use for the evaluation of system effectiveness in INEX are based on certain assumptions of user behaviour which are not empirically validated. The track also aimed to investigate those assumptions.

Four topics from the ad hoc CO topic set were reused and classified into two topic types: background and comparison. Experiments were designed to isolate the effect of topic and searcher from that of search system so that to investigate searcher behaviours. The narratives, which included a work task scenario, provided searchers (users) with background information needs. An online baseline system was provided to the participants. Ten institutions participated with the fulfilment of a minimum requirement of 8 searchers participating in the experiments. Data gathering sources were questionnaires, informal interviews and log files. Initial analysis and findings are presented in [7].

10 Conclusion and Outlook

INEX 2004 was a success and showed that XML retrieval is a challenging field within IR and related research. In addition to learning more about XML retrieval approaches, INEX 2004 has also made further steps in the evaluation methodology for XML retrieval.

INEX 2005 will start in April of this year, and in addition to its current five tracks, will have two new tracks: XML multimedia track and document mining track. The former will look at issues regarding the access to multimedia content embedded in XML document; and the latter is concerned with clustering and categorising tasks in the context of XML documents.

References

1. N. Fuhr, N. Gövert, and K. Großjohann. HyREX: Hypermedia retrieval engine for XML. In *Proceedings of the 25th Annual International Conference on Research and Development in Information Retrieval*, page 449, 2002. Demonstration.
2. N. Fuhr, N. Gövert, G. Kazai, and M. Lalmas, editors. *INitiative for the Evaluation of XML Retrieval (INEX). Proceedings of the First INEX Workshop. Dagstuhl, Germany, December 8–11, 2002*, ERCIM Workshop Proceedings, March 2003. <http://www.ercim.org/publication/ws-proceedings/INEX2002.pdf>.

3. N. Gövert and G. Kazai. Overview of the INitiative for the Evaluation of XML retrieval (INEX) 2002. In Fuhr et al. [2], pages 1–17. <http://www.ercim.org/publication/ws-proceedings/INEX2002.pdf>.
4. N. Gövert, G. Kazai, N. Fuhr, and M. Lalmas. Evaluating the effectiveness of content-oriented XML retrieval. Technical report, University of Dortmund, Computer Science 6, 2003. http://www.is.informatik.uni-duisburg.de/bib/pdf/ir/Goevert_etal:03a.pdf.
5. V. V. Raghavan, P. Bollmann, and G. S. Jung. Retrieval system evaluation using recall and precision: Problems and answers. In *Proceedings of the Twelfth Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 59–68, 1989.
6. Z. Szlávik, and T. Rölleke. Building and Experimenting with a Heterogeneous Collection. In this volume, 2005.
7. A. Tombros, B. Larsen, and S. Malik. The Interactive Track at INEX 2004 In this volume, 2005.
8. A. Trotman, and B. Sigurbjörnsson. Narrowed Extended XPath I (NEXI). In this volume, 2005.