

Overview of the INitiative for the Evaluation of XML retrieval (INEX) 2002

Norbert Gövert
University of Dortmund
Germany
goevert@ls6.cs.uni-dortmund.de

Gabriella Kazai
Queen Mary University of London
United Kingdom
gabs@dcs.qmul.ac.uk

The INitiative for the Evaluation of XML retrieval (INEX) aims at providing an infrastructure for evaluating the effectiveness of content-oriented XML retrieval. In the first round of INEX, in 2002, a test collection of real world XML documents along with standard topics and respective relevance assessments has been created. Research groups from 36 different organisations participated in this collaborative effort. In this article we describe the test collection and how it was constructed. An overview of the metrics used to evaluate the effectiveness of XML retrieval approaches and of the evaluation results of 51 submissions from the INEX 2002 participants is also provided.

1 Introduction

The INitiative for the Evaluation of XML retrieval (INEX) was set up at the beginning of 2002 with the aim to establish an infrastructure and to provide means, in the form of a large XML test collection and appropriate scoring methods, for the evaluation of content-oriented retrieval of XML documents. INEX 2002 was the first in a series of future XML retrieval evaluation efforts. As a result of a collaborative effort, during the course of 2002, INEX created an XML test collection consisting of publications of the IEEE Computer Society between 1995 and 2002, 60 topics, and graded relevance assessments. Using the constructed test collection and the developed set of evaluation metrics and procedures, the retrieval effectiveness of the participating organisations' XML retrieval approaches were evaluated and their results compared.

This paper presents an overview of INEX 2002, the constructed test collection and the developed evaluation metrics, and provides a summary of the research in XML retrieval described in detail in the remainder of the proceedings. Although this overview is intended to provide a complete account of INEX 2002, it does not aim to explain or review the underlying research concepts for the evaluation of XML retrieval. On the other hand, for completeness, we cover in this paper some material already published at the SIGIR XML Workshop in 2002 while the initiative was still in progress and which provided an introduction into INEX [2].

The paper is structured as follows. In Section 2 we provide a brief summary of the INEX participants and their systems. Section 3 outlines the evaluation task set by INEX. Section 4 provides an overview of the INEX test collection along with a description of how the collection was constructed. In Section 5 a specification of the evaluation metrics applied for INEX 2002 is given, and Section 6 summarises the evaluation results. We end with conclusions and an outlook on INEX 2003 in Section 7.

2 Participating organisations

In response to the call for participation, issued in March 2002, 49 organisations from 21 countries on four continents registered within six weeks. However, throughout the year a number of groups dropped out due to resource

requirements, while a number of new groups joined the initiative at the relevance assessments stage. The final 36 active INEX 2002 groups are listed in Table 1.

Due to the diversity in the background of the participating groups, a wide range of different approaches to XML retrieval were represented within INEX 2002. Although the approaches are quite diverse, we tried to classify them using the following three categories [2]:

IR model-oriented: Research groups that focus on the extension of a specific type of information retrieval (IR) model (e. g. vector space, rule-based, logistic regression, LSI), which they have applied to standard IR test collections in the past, to deal with XML documents.

DB-oriented: Groups that are working on extending database (DB) management systems to deal with semistructured data; most of these groups also incorporate uncertainty weights, thus producing ranked results.

XML-specific: Groups that, instead of aiming to extend existing approaches towards XML, have developed models and systems specifically for XML. Although these groups have very different backgrounds they usually base their work on XML standards (like XSL, XPath or XQuery).

Table 1 shows the approaches followed by the different groups. As it can be seen, most of the retrieval approaches were pure IR, DB or XML, although a few groups combined elements from two categories.

3 The task

Evaluation initiatives for flat document retrieval in IR, such as TREC¹, include several different tracks focusing on tasks such as ad-hoc retrieval, routing, filtering, and interactive retrieval, etc. Although most of these tasks are applicable to XML document retrieval, this being the first year of the initiative, we decided to run only one track, where the task to be performed was set as the ad-hoc retrieval of XML documents. Just as in TREC, the ad-hoc task was defined with the aim to evaluate the performance of systems that search a static set of documents using a new set of topics. This task has been described as a simulation of how a library might be used, where the collection of documents is known, while the queries to be asked are unknown [13]. Compared with flat document retrieval, however, for the evaluation of the ad-hoc retrieval of XML documents, we needed to consider additional requirements.

Given the different approaches to XML document retrieval (Section 2) and the widespread development and use of XML query languages, users of XML retrieval systems are able to issue (directly or indirectly) more complex queries than those used in flat document retrieval. For example, users are able to exploit the structural nature of the data and restrict their search to specific structural elements within an XML collection. This has to be reflected in the queries used for the evaluation of such systems. Content-oriented XML retrieval systems, however, should also support queries that do not specify structural conditions. The need for this type of queries for the evaluation of XML retrieval is well published (even within this proceedings) and stems from the fact that users often do not know the exact structure of the XML documents. Taking this into account, we identified the following two types of queries to be included in the INEX ad-hoc task:

Content-and-structure (CAS) queries are topic statements that contain explicit references to the XML structure, either by restricting the context of interest or the context of certain search concepts.

Content-only (CO) queries ignore the document structure and are, in a sense, the traditional topics used in IR test collections. Their resemblance to traditional IR queries is, however, only in their appearance. They pose a challenge to XML retrieval in that the retrieval results to such queries can be (possibly overlapping) XML elements of varying granularity that fulfill the query.

The objective of the evaluation in INEX, based on the ad-hoc task, is to assess a system's retrieval effectiveness, where effectiveness is measured as a system's ability to satisfy both content and structural aspects of a user's information need and retrieve the most specific relevant document components, which are exhaustive to the topic of request and match its structural constraints.

¹<http://trec.nist.org/>

Organisation	Retrieval approach	no of runs submitted	Assessed topics
Carnegie Mellon University	IR		07, 28
Centrum voor Wiskunde en Informatica (CWI)	DB+IR	3	02, 03, 36
CSIRO Mathematical and Information Sciences	IR	3	14, 15, 27
doctronic GmbH	IR+XML	1	43
Electronics and Telecommunications Research Institute (ETRI)	DB+XML	1	26, 58
ETH Zurich	DB+IR	1	16, 47
Florida A&M University			59
IBM Haifa Labs	IR	3	08, 09
Institut de Recherche en Informatique de Toulouse (IRIT)	IR	1	
Nara Institute of Science and Technology	IR	1	37, 38
Queen Mary University of London	IR	3	53
Queensland University of Technology	IR+XML	3	29, 60
Royal School of Library and Information Science	other	3	04, 34
Salzburg Research Forschungsgesellschaft	IR	1	
Sejong Cyber University	XML	1	25
Tarragon Consulting Corporation	IR	2	31, 33
Universität Bayreuth	DB	1	05, 06
Universität Dortmund / Universität Duisburg-Essen	IR	3	30
Université Pierre et Marie Curie	IR+XML	3	10, 45, 50
University of Amsterdam	IR	3	01, 42
University of California, Berkeley	IR	3	17, 18
University of California, Los Angeles		1	48, 49
University of Helsinki	IR		19, 51
University of Melbourne	IR	3	20, 52
University of Michigan	DB+XML	2	12, 13
University of Minnesota Duluth	IR	1	11, 46
University of North Carolina at Chapel Hill	IR	1	
University of Rostock	XML		21, 22
University of Twente	DB	3	23, 24
University of Zurich			41
Organisations joined at the relevance assessments stage:			
Dublin City University			39, 40
Ecole Nationale Supérieure des Mines de Saint-Etienne			50
Justus-Liebig-Universität Gießen			50
University of California, San Diego			32
University of East Anglia			40
University of Granada			44

Table 1: List of INEX 2002 participants

id	Publication title	Year	Size (MB)	no of articles
an	IEEE Annals of the History of Computing	1995-2001	13.2	316
cg	IEEE Computer Graphics and Applications	1995-2001	19.1	680
co	Computer	1995-2001	40.4	1 902
cs	IEEE Computational Science & Engineering	1995-1998	14.6	571
	Computing in Science & Engineering	1999-2001		
dt	IEEE Design & Test of Computers	1995-2001	13.6	539
ex	IEEE Expert	1995-1997	20.3	702
	IEEE Intelligent Systems	1998-2001		
ic	IEEE Internet Computing	1997-2001	12.2	547
it	IT Professional	1999-2001	4.7	249
mi	IEEE Micro	1995-2001	15.8	604
mu	IEEE MultiMedia	1995-2001	11.3	465
pd	IEEE Parallel & Distributed Technology	1995-1996	10.7	363
	IEEE Concurrency	1997-2000		
so	IEEE Software	1995-2001	20.9	936
tc	IEEE Transactions on Computers	1995-2002	66.1	1 042
td	IEEE Transactions on Parallel & Distributed Systems	1995-2002	58.8	765
tg	IEEE Transactions on Visualization & Computer Graphics	1995-2002	15.2	225
tk	IEEE Transactions on Knowledge and Data Engineering	1995-2002	48.1	585
tp	IEEE Transactions on Pattern Analysis & Machine Intelligence	1995-2002	62.9	1 046
ts	IEEE Transactions on Software Engineering	1995-2002	46.1	570
Total			494	12 107

Table 2: The INEX document collection

4 The test collection

Similarly to standard IR test collections, the INEX test collection consists of three parts: a set of documents, topics and relevance assessments.

4.1 Documents

The document collection was donated to INEX by the IEEE Computer Society. It consists of the fulltexts 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. Table 2 lists some statistics for the different publications included in the collection. Although the size of the document collection is relatively small compared with TREC, it has a suitably complex XML structure containing 192 different content models in its DTD. On average, an article contains 1 532 XML nodes, where the average depth of a node is 6.9.

All documents in the collection are tagged using XML conforming to one common schema, i. e. DTD. Figure 1 shows the overall structure of a typical article consisting of a front matter (<fm>), a body (<body>), and a back matter (<bm>). The front matter contains the article's metadata, such as title, author, publication information, and abstract. Following it is the article's body, which contains the content. The body is structured into sections (<sec>), sub-sections (<ss1>), and sub-sub-sections (<ss2>). These logical units start with a title, followed by a number of paragraphs. In addition, the content has markup for references (citations, tables, figures), item lists, layout (such as emphasised and bold faced text), etc. The back matter contains a bibliography and information about the authors of the article.

4.2 Topics

The topic format and the topic development procedures were based on TREC guidelines, which were modified to accommodate the two types of topics used: CO and CAS (see Section 3).

```

<article>
  <fm>
    ...
    <ti>IEEE Transactions on ...</ti>
    <atl>Construction of ...</atl>
    <au>
      <fnm>John</fnm>
      <snm>Smith</snm>
      <aff>University of ...</aff>
    </au>
    <au>...</au>
    ...
  </fm>
  <bdy>
    <sec>
      <st>Introduction</st>
      <p>...</p>
      ...
    </sec>
    ...
  </bdy>
  <sec>
    <st>...</st>
    ...
    <ssl>...</ssl>
    <ssl>...</ssl>
    ...
  </sec>
  ...
</bdy>
<bm>
  <bib>
    <bb>
      <au>...</au><ti>...</ti>
      ...
    </bb>
    ...
  </bib>
</bm>
</article>

```

Figure 1: Sketch of the structure of the typical INEX articles

```

<!ELEMENT INEX-Topic (Title, Description, Narrative, Keywords)>
<!ATTLIST INEX-Topic
  topic-id CDATA #REQUIRED
  query-type CDATA #REQUIRED
  ct-no CDATA #REQUIRED
>
<!ELEMENT Title ( te?, (cw, ce?)+ )>
<!ELEMENT te (#PCDATA)>
<!ELEMENT cw (#PCDATA)>
<!ELEMENT ce (#PCDATA)>
<!ELEMENT Description (#PCDATA)>
<!ELEMENT Narrative (#PCDATA)>
<!ELEMENT Keywords (#PCDATA)>

```

Figure 2: Topic DTD

4.2.1 Topic format

The topic format was modified to allow the definition of containment conditions and the specification of target elements (e. g. elements that should be returned to the user). The DTD of an INEX topic is shown in Figure 2. The four main parts of a topic are the topic title, topic description, narrative and keywords.

As in TREC, the topic title is a short version of the topic description and usually consists of a number of keywords that best describe what the user is looking for. In INEX, however, the topic title serves as a summary of both content and structure related requirements of a user's information need. An INEX topic title, hence, may contain a number of different components: target elements (<te>), a set of search concepts (<cw>), and a set of context elements (<ce>). The combination of the latter two corresponds to a containment condition. A search concept may be represented by a set of keywords or phrases. A CO topic title consists only of <cw> components as, by definition, it does not specify constraints over the structure of the result elements. For CAS queries, a topic title may specify the target elements of the search and/or the context elements of given search concepts. Both target and context elements may list one or more XML elements (e. g. <ce>abs, kwd</ce>), which may be given by their absolute (e. g. article/fm/au) or abbreviated path (e. g. //au), or by their element type (e. g. au). Omitting the target or context element in a topic title indicates that there are no restrictions placed upon the type of element the search should return, or the type of element a given concept should be a subject of.

The topic description is a one- or two-sentence natural language definition of the information need. The narrative is a detailed explanation of the topic statement and a description of what makes a document/component relevant or

```

<INEX-Topic topic-id="09" query-type="CAS" ct-no="048">
  <Title>
    <te>article</te>
    <cw>non-monotonic reasoning</cw> <ce>bdy/sec</ce>
    <cw>1999 2000</cw> <ce>hdr//yr</ce>
    <cw>-calendar</cw> <ce>tig/at1</ce>
    <cw>belief revision</cw>
  </Title>
  <Description>
    Retrieve all articles from the years 1999-2000 that deal with works on non-
    monotonic reasoning. Do not retrieve articles that are calendar/call for papers.
  </Description>
  <Narrative>
    Retrieve all articles from the years 1999-2000 that deal with works on non-
    monotonic reasoning. Do not retrieve articles that are calendar/call for papers.
  </Narrative>
  <Keywords>
    non-monotonic reasoning belief revision
  </Keywords>
</INEX-Topic>

```

Figure 3: A CAS topic from the INEX test collection

```

<INEX-Topic topic-id="45" query-type="CO" ct-no="056">
  <Title>
    <cw>augmented reality and medicine</cw>
  </Title>
  <Description>
    How virtual (or augmented) reality can contribute to improve the medical and
    surgical practice.
  </Description>
  <Narrative>
    In order to be considered relevant, a document/component must include
    considerations about applications of computer graphics and especially augmented
    (or virtual) reality to medicine (including surgery).
  </Narrative>
  <Keywords>
    augmented virtual reality medicine surgery improve computer assisted aided image
  </Keywords>
</INEX-Topic>

```

Figure 4: A CO topic from the INEX test collection

not. The keywords component of a topic was added in INEX as a means to keep a record of the list of search terms used for retrieval during the topic development process carried out by the participating groups (see Section 4.2.2).

The three attributes of a topic are: `topic-id` (e. g. 1 to 60), `query-type` (e. g. CAS or CO), and `ct-no`, which refers to the candidate topic number (e. g. 1 to 143). Figures 3 and 4 show examples for both types of topics.

4.2.2 The topic development process

In INEX, the topics were created by the participating groups. We asked each organisation to create a set of candidate topics that were representative of what real users might ask and the type of the service that operational systems may provide. Participants were provided with guidelines to assist them in this task [5]. The guide identified the following stages of the topic creation process: (1) Creation of the initial topic statement, (2) Collection exploration, (3) Topic refinement, and (4) Topic selection. While the first three stages were carried out by the participants, the selection of the final topics was left to us.

During the first stage participants created their initial topic statements. 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-biased queries.

	CAS	CO
no of topics	30	30
total no of <cw> components	62	30
avg no of <cw> / topic title	2.06	1.0
avg no of unique words / cw	2.5	4.3
avg no of unique words / topic title	5.1	4.3
total no of <ce> components	49	0
avg no of <ce> / topic title	1.63	–
avg no of XML elements / <ce>	1.53	–
avg no of XML elements / topic title	2.5	–
no of topics with <ce> representing a fact	12	–
no of topics with <ce> representing content	6	–
no of topics with mixed <ce>	12	–
total no of topics with <te> components	25	0
avg no of XML elements / <te>	1.68	–
no of topics with <te> representing a fact	13	–
no of topics with <te> representing content	12	–
no of topics with <te> representing articles	6	–
total no of (<cw>, <ce>) pairs	49	0
avg no of (<cw>, <ce>) pairs / topic title	1.63	–
avg no of words in topic description	18.8	16.1
avg no of words in keywords component	7.06	8.7

Table 3: Statistics on CAS and CO queries in the INEX test collection

During the collection exploration stage, participants estimated the number of relevant documents/components to their candidate topics. Unlike TREC, we did not provide topic authors a retrieval system for this task, but participants used their own retrieval engines. They then judged the top 25 retrieved components and the top 100 results after performing relevance feedback. Keywords used in the retrieval runs were recorded within the topic’s keywords component.

In the topic refinement stage the components of a topic were finalised ensuring coherency and that each component could be used in a stand-alone fashion (e. g. retrieval using only the topic title).

After completion of the first three stages, the candidate topics were submitted to INEX. A total of 143 candidate topics were received, of which 60 topics (30 CAS and 30 CO) were selected into the final set of topics. The selection of the final 60 topics was based on the combination of criteria, such as including equal number of CO and CAS topics, having topics that are representative of IR, DB and XML-specific search situations, balancing the load across participants for relevance assessments, and eliminating topics that were considered too ambiguous or too difficult to judge. We also aimed to include topics that were likely to retrieve diverse sets (varying granularity) of relevant components. Furthermore, we based topic selection on the estimated number of relevant components, where we selected topics with at least 2, but no more than 20 relevant items in the top 25 retrieved components. Note that due to the lack of information with respect to the estimated number of relevant components within the top 100 results after relevance feedback, this data was largely ignored during topic selection.

Table 3 shows some statistics on the final set of INEX topics. Note that these figures are different from that in [2] as a result of subsequent changes to the topics. In the statistics we differentiated between context and target elements that represent facts, such as author or title information, or content, such as the text of an article or a part of the article. Looking at the 25 CAS topics that specified target elements, we can see that more than half requested facts to be returned to the user. Furthermore, the majority of the CAS topics contained either only fact (e. g. specifying the publication year and/or the title), or a mixture of fact and content containment conditions (e. g. specifying the author and the subject of a document component).

	CAS topics	CO topics
no of documents submitted	64 024	97 947
no of documents in pools	23 375	30 275
reduction	63 %	69 %
no of components submitted	100 904	139 235
no of components in pools	47 419	60 066
reduction	53 %	57 %

Table 4: Pooling effect for CAS and CO topics

4.3 Submissions

Participating groups evaluated the final set of topics against the document collection and produced, for each topic, a ranked list of XML documents / components (result elements). The top 100 result elements from all sixty sets of ranked lists (one per topic) consisted the results of one retrieval run. Each group was allowed to submit up to three runs. The submission format and procedure is detailed in [7]. Each result element was identified using a combination of file names and XPath. The file name and file path uniquely identified an article within the document collection, and XPath allowed the location of a given component within the XML tree of the article. The result components varied from author, title and paragraph elements through sub-section and section elements to complete articles and even journals. Associated with a result element were its retrieval rank and/or its relevance status value.

In the first round of INEX, a total of 51 runs were submitted by 25 participating organisations. 42 of the 51 submissions contained results for the CAS topics and 49 contained results for the CO topics.

For each topic, all of the results from the submissions were merged to form the pool for assessment [11]. A median sized assessment pool for CAS topics contained 1 585 document components from 749 different articles. For CO topics the median sized assessment pool contained 1 980 document components from 981 different articles. Table 4 shows the pooling effect for CAS and CO topics.

4.4 Assessments

The assessment pools were then assigned to participants for assessment; either to the original topic authors or when this was not possible, on a voluntary basis, to groups with expertise in the topic's subject area. The topics assessed by the different groups are summarised in Table 1. Note that the list excludes topics 35, 54, 55, 56, and 57 as no groups volunteered to assess them. On the other hand, we obtained multiple assessments for topics 40 and 50, which were assessed by two and three assessors, respectively. We will analyse these sets in the near future to estimate the consistency of the collected assessments.

The assessments were done along the following two dimensions:

Topical relevance, which reflects the extent to which the information contained in a document component satisfies the information need.

Component coverage, which reflects the extent to which a document component is focused on the information need, while being an informative unit.

Both these dimensions were measured using graded scales. For topical relevance we used the following four-point scale [8]:

Irrelevant (0): The document component does not contain any information about the topic of request.

Marginally relevant (1): The document component mentions the topic of request, but only in passing.

Fairly relevant (2): The document component contains more information than the topic description, but this information is not exhaustive. In the case of multi-faceted topics, only some of the sub-themes or viewpoints are discussed.

Highly relevant (3): The document component discusses the topic of request exhaustively. In the case of multi-faceted topics, all or most sub-themes or viewpoints are discussed.

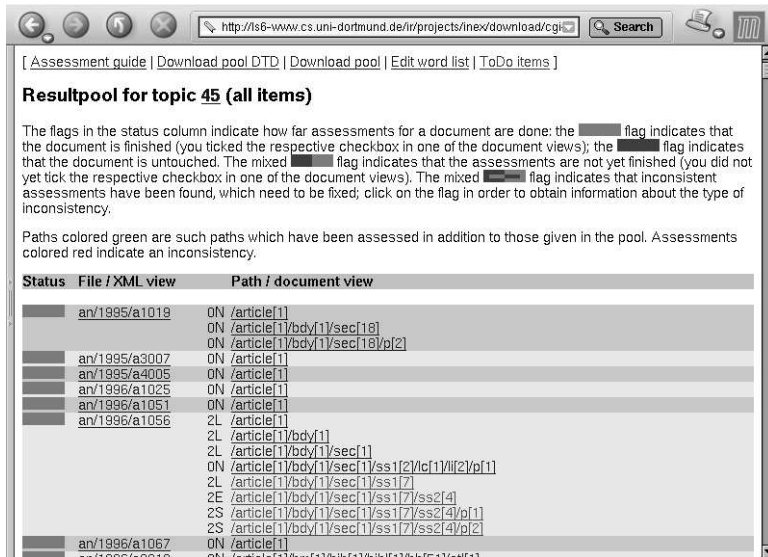


Figure 5: Result pool. Result elements are listed in alphabetical order and grouped within article elements. The relevance and coverage values are shown in front of assessed elements.

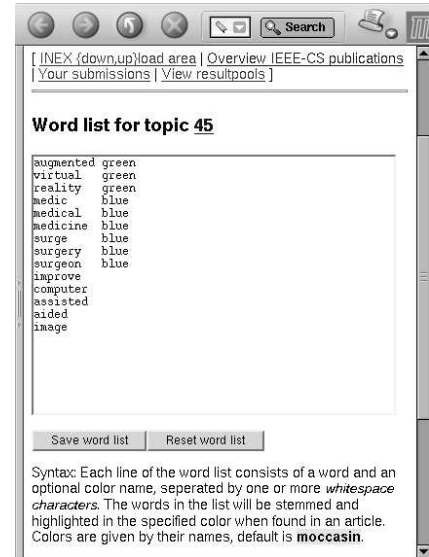


Figure 6: Word list editor. It was used by the assessors to specify a list of cue terms that were then highlighted in the document views.

Component coverage was selected from the following four categories [10]:

No coverage (N): The topic or an aspect of the topic is not a theme of the document component.

Too large (L): The topic or an aspect of the topic is only a minor theme of the document component.

Too small (S): The topic or an aspect of the topic is the main or only theme of the document component, but the component is too small to act as a meaningful unit of information.

Exact coverage (E): The topic or an aspect of the topic is the main or only theme of the document component, and the component acts as a meaningful unit of information.

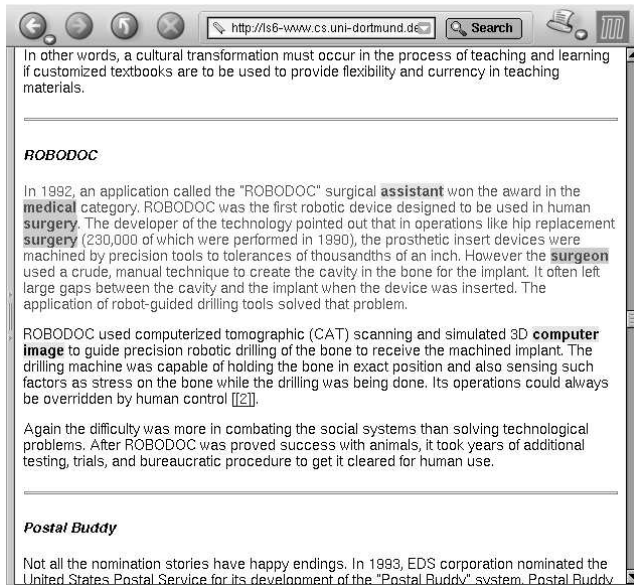
Note that the two assessed dimensions are not perfectly orthogonal to each other. Some combinations of relevance/coverage values do not make sense: A component which has no relevance cannot have any coverage with the topic. Vice versa, if a document component has no coverage with a topic, it cannot be relevant to the topic at the same time. In a similar way, a document component which has a coverage too small, cannot be highly relevant, since this would assume that all or most of the concepts requested by the topic are discussed exhaustively.

Assessors were sent detailed instructions on how to carry out the assessments based on the above two dimensions [6]. Assessments were recorded using an on-line assessment system, which allowed users to view the pooled result set of a given topic, to browse the document collection and view articles and result elements both in XML (i. e. showing the tags) and document view (i. e. formatted for ease of reading). Other features included facilities such as keyword highlighting, and consistency checking of the assessments. Figures 5, 6, and 7 show screenshots of the assessment system.

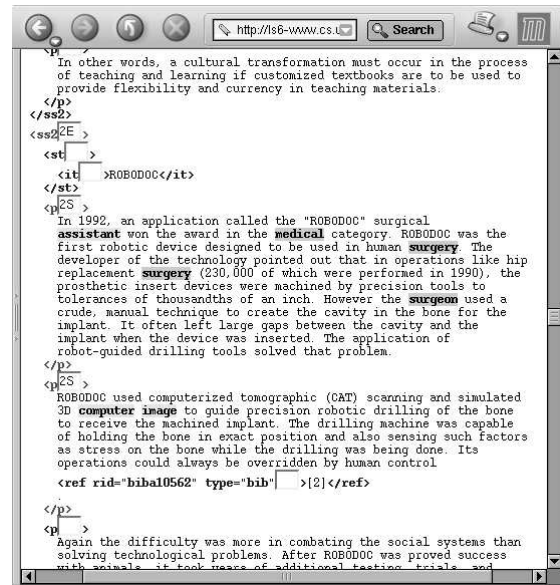
Table 5 shows a summary of the collected assessments for CAS and CO topics. Here, the relatively large proportion of non-article level elements with exact coverage compared with article elements indicates that for most topics sub-components were considered as the preferred units to be returned to a user; this is emphasised in Figure 8. Figure 9 shows the relative distribution of selected non-article XML elements that were judged relevant.

5 Evaluation metrics

Due to the nature of XML retrieval, metrics from traditional evaluation initiatives like TREC and CLEF could not be applied in INEX without modification. Therefore, it was necessary to develop new evaluation procedures. Here we



a) Document view



b) XML view

Figure 7: A section of an article in document and XML view. Result elements are highlighted and cue words are marked as specified in the word list editor. Participants used the XML view to record their assessments, i. e. values of relevance and coverage for a given XML element.

Rel+ Cov	CAS topics		CO topics	
	article level	non-articles	article level	non-articles
3E	187	2 304	307	1 087
2E	59	1 128	165	1 107
1E	82	1 770	114	827
3L	173	424	394	1 145
2L	137	507	599	2 295
1L	236	719	854	2 708
2S	21	846	118	3 825
1S	54	1 119	116	3 156
All	949	8 817	2 667	16 150

Table 5: Assessments at article and component levels

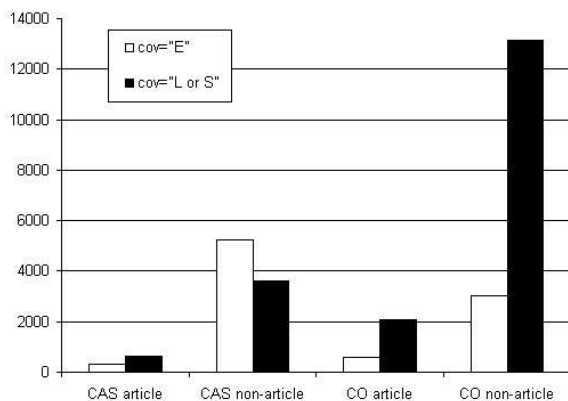


Figure 8: Distribution of relevant article and non-article elements (topical relevance > 0).

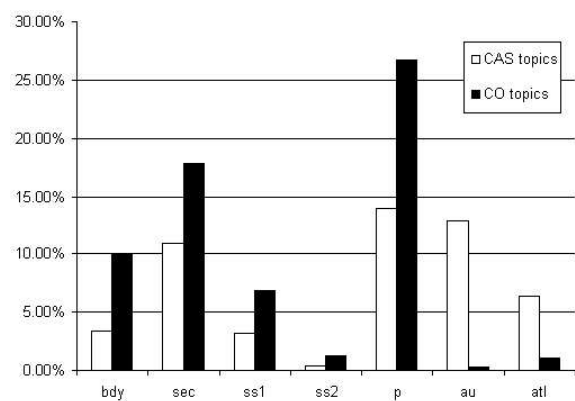


Figure 9: Distribution of relevant non-article elements (topical relevance > 0).

describe the evaluation metrics that were discussed at the INEX Workshop and have been applied to the INEX 2002 submissions. These metrics have been implemented within the `inex_eval` package, which has been distributed to the participants. In addition, a Web-based evaluation interface has also been provided for the participants.

In Section 5.1 we describe how implicit assessments have been derived from the explicit assessments done by the assessors. The evaluation metrics proposed in Section 5.3 are based on established recall/precision metrics. However, in order to apply these in INEX, the two dimensional quality assessments (see Section 4.4) first had to be quantised onto a binary relevance scale. The quantisation functions developed for this purpose are given in Section 5.2.

5.1 Implicit relevance assessments

Due to the nature of the two assessed dimensions (*topical relevance* and *component coverage*) and from the INEX quality assessment guide [6] one can, in certain cases, deduce assessments for nodes which have not been assessed explicitly:

- Due to the definition of the relevance dimension, the relevance level of a parent component of an assessed component is equal to or greater than the relevance of the assessed component.
- For a component which has a coverage assessment of *exact* or *too large* it can be deduced that its parent component has a coverage of *too large*.

These rules have been applied recursively, up to the article level of the documents, in order to add implicit assessments to the explicit assessments done by the assessors. The only exception for applying the rules are CAS topics with *target element* specifications, as it has been agreed to interpret the target element specifications in a strict way in terms of evaluation.

5.2 Quantisation of relevance and coverage

In order to apply traditional recall/precision metrics, values for the two dimensions of relevance and coverage must be quantised by some function f_{quant} to a single relevance value:

$$\begin{aligned} f_{quant} &: Relevance \times Coverage \rightarrow [0, 1] \\ &(rel, cov) \mapsto f_{quant}(rel, cov) \end{aligned} \quad (1)$$

Here, the set of relevance assessments is $Relevance := \{0, 1, 2, 3\}$, and the set of coverage assessments is $Coverage := \{N, S, L, E\}$.

Quantisation functions can be selected according to the desired user standpoint. For INEX 2002, two different functions have been selected: f_{strict} and $f_{generalised}$.

The quantisation function f_{strict} is used to evaluate whether a given retrieval method is capable of retrieving highly relevant and highly focused document components:

$$f_{strict}(rel, cov) := \begin{cases} 1 & \text{if } rel = 3 \text{ and } cov = E, \\ 0 & \text{else} \end{cases} \quad (2)$$

Other functions can be based on the different possible combinations of relevance degrees and coverage categories, such as $f_{quant}(rel, cov) = 1$ if $rel > 1$ and $cov = E$. In order to credit document components according to their *degree of relevance* (generalised recall/precision), the quantisation function $f_{generalised}$ is used:

$$f_{generalised}(rel, cov) := \begin{cases} 1.00 & \text{if } (rel, cov) = 3E, \\ 0.75 & \text{if } (rel, cov) \in \{2E, 3L\}, \\ 0.50 & \text{if } (rel, cov) \in \{1E, 2L, 2S\}, \\ 0.25 & \text{if } (rel, cov) \in \{1S, 1L\}, \\ 0.00 & \text{if } (rel, cov) = 0N \end{cases} \quad (3)$$

5.3 Recall / precision metrics

Given the type of quantisation described above, each document component in a result ranking is assigned a single relevance value. In INEX 2002, overlaps of document components in rankings were ignored, thus procedures that calculate recall/precision curves for standard document retrieval could be applied directly to the results of the quantisation functions. The method described by Raghavan et al. in [9] is used for this. Here, precision is interpreted as the probability, $P(\text{rel}|\text{retr})$, that a document viewed by a user is relevant. Given that the user stops viewing at the ranking after a given number of relevant document components NR , this probability can be computed as:

$$P(\text{rel}|\text{retr})(NR) := \frac{NR}{NR + esl_{NR}} = \frac{NR}{NR + j + s \cdot i / (r + 1)}. \quad (4)$$

The expected search length, esl_{NR} , denotes the total number of non-relevant document components that are estimated to be retrieved until the NR th relevant document is retrieved. Let l denote the rank from which the NR th relevant component is drawn. Then j is the number of non-relevant document components within the ranks before rank l , s is the number of relevant components to be taken from rank l , and r and i are the numbers of relevant and non-relevant components in rank l , respectively (details on the derivation are given by Cooper in [1]).

Raghavan et al. also gave theoretical justification, that intermediary real numbers can be used instead of simple recall points only (here, n is the total number of relevant document components with regard to the user request in the collection; $x \in [0, 1]$ denotes an arbitrary recall value):

$$P(\text{rel}|\text{retr})(x) := \frac{x \cdot n}{x \cdot n + esl_{x \cdot n}} = \frac{x \cdot n}{x \cdot n + j + s \cdot i / (r + 1)}. \quad (5)$$

This leads to an intuitive method for employing arbitrary fractional numbers, x , as recall values and thus allows for averaging evaluation results over multiple topic results.

The metric from Raghavan et al. has some theoretical advantages over the metric described in [12]: besides the intuitive method for interpolation it handles weakly ordered ranks correctly. The main advantage, however, is that the variables n , j , i , r , and s in Formula 5 can be interpreted as expectations, thus allowing for a straightforward implementation of the metric for the generalised quantisation function. For example, given a function $\text{assessment}(c)$, which yields the relevance/coverage assessment for a given document component c , the number n of relevant components with respect to a given topic and quantisation function is computed as:

$$n = \sum_{c \in \text{components}} \mathbf{f}_{\text{quant}}(\text{assessment}(c)). \quad (6)$$

Expectations for the other variables are computed respectively. Table 6 lists the number of relevant document components on a per topic basis, for both quantisation functions $\mathbf{f}_{\text{strict}}$ and $\mathbf{f}_{\text{generalised}}$.

For computation of the recall/precision curves for a given submission using Raghavan et al.'s method, it is assumed that the submission conceptually ranks all components available through the document collection. In INEX 2002, however, participants were allowed to submit 100 document components per topic only. The evaluation procedure therefore creates a virtual final rank, which enumerates all the components not being part of the set of components explicitly ranked within the submission itself. A theoretical problem which arises in the case of structured document retrieval is the question of the size of this rank (needs to be determined in order to apply Formula 5). Obviously, not every element given by the XML markup of the documents are candidates for retrievable components (most of them would be far too small to serve as a meaningful unit of information). We therefore computed a rough estimation of this figure, based on the assessments available for a given topic. For this, it is assumed that for documents where explicit assessments are available, *all retrievable* components have been assessed (explicitly or implicitly). In addition, it is assumed that retrievable components are distributed equally in all documents, regardless of the fact whether they have been assessed or not. The estimated number of retrievable components for a given topic can then be computed by:

$$|\text{components}| \approx |\text{documents}| \cdot \frac{|\text{components assessed}|}{|\text{documents assessed}|} \quad (7)$$

The number of components per topic in Table 6 have been computed this way.

	strict		generalised			strict		generalised	
	comp.	rel.	comp.	rel.		comp.	rel.	comp.	rel.
01	14 222	44.00	14 222	44.00	31	15 366	4.00	15 366	45.25
02	12 160	567.00	12 160	577.50	32	141 858	35.00	141 858	795.50
03	48 360	125.00	48 360	831.50	33	13 235	2.00	13 235	34.50
04	26 535	41.00	26 535	105.00	34	26 336	66.00	26 336	412.50
05	14 373	79.00	14 373	126.50	35	–	–	–	–
06	12 186	17.00	12 186	91.25	36	17 507	31.00	17 507	138.75
07	35 246	55.00	35 246	174.50	37	42 102	138.00	42 102	860.50
08	12 220	8.00	12 220	9.00	38	48 006	111.00	48 006	1 304.00
09	12 107	10.00	12 107	10.25	39	105 503	48.00	105 503	277.25
10	30 237	57.00	30 237	272.50	40	13 587	124.00	13 587	232.50
11	15 703	73.00	15 703	252.00	41	22 691	57.00	22 691	159.00
12	22 191	30.00	22 191	57.50	42	63 129	91.00	63 129	309.50
13	19 109	1.00	19 109	2.75	43	49 528	15.00	49 528	77.75
14	72 339	30.00	72 339	172.00	44	65 139	36.00	65 139	158.00
15	90 572	39.00	90 572	690.25	45	31 845	57.00	31 845	535.75
16	12 107	91.00	12 107	122.25	46	19 962	26.00	19 962	239.50
17	97 025	21.00	97 025	78.25	47	78 780	22.00	78 780	233.75
18	30 690	7.00	30 690	66.25	48	21 349	65.00	21 349	296.75
19	15 392	71.00	15 392	152.25	49	21 792	9.00	21 792	157.25
20	149 009	33.00	149 009	83.50	50	133 437	0.00	133 437	451.50
21	45 082	9.00	45 082	114.50	51	15 548	26.00	15 548	191.25
22	29 436	73.00	29 436	95.75	52	135 699	15.00	135 699	140.50
23	14 562	29.00	14 562	36.75	53	76 783	34.00	76 783	816.25
24	12 107	6.00	12 107	12.25	54	–	–	–	–
25	15 303	8.00	15 303	24.50	55	–	–	–	–
26	15 948	174.00	15 948	280.50	56	–	–	–	–
27	1 809 996	149.00	1 809 996	149.00	57	–	–	–	–
28	12 107	47.00	12 107	47.00	58	28 576	210.00	28 576	722.75
29	33 703	173.00	33 703	618.00	59	–	–	–	–
30	47 453	424.00	47 453	758.25	60	26 318	174.00	26 318	638.50

a) CAS topics

b) CO topics

Table 6: Number of components (comp.) and relevant components (rel.) per topic, for both quantisation functions. The number of relevant components has been computed using Equation 6, while the number of components has been estimated using Equation 7.

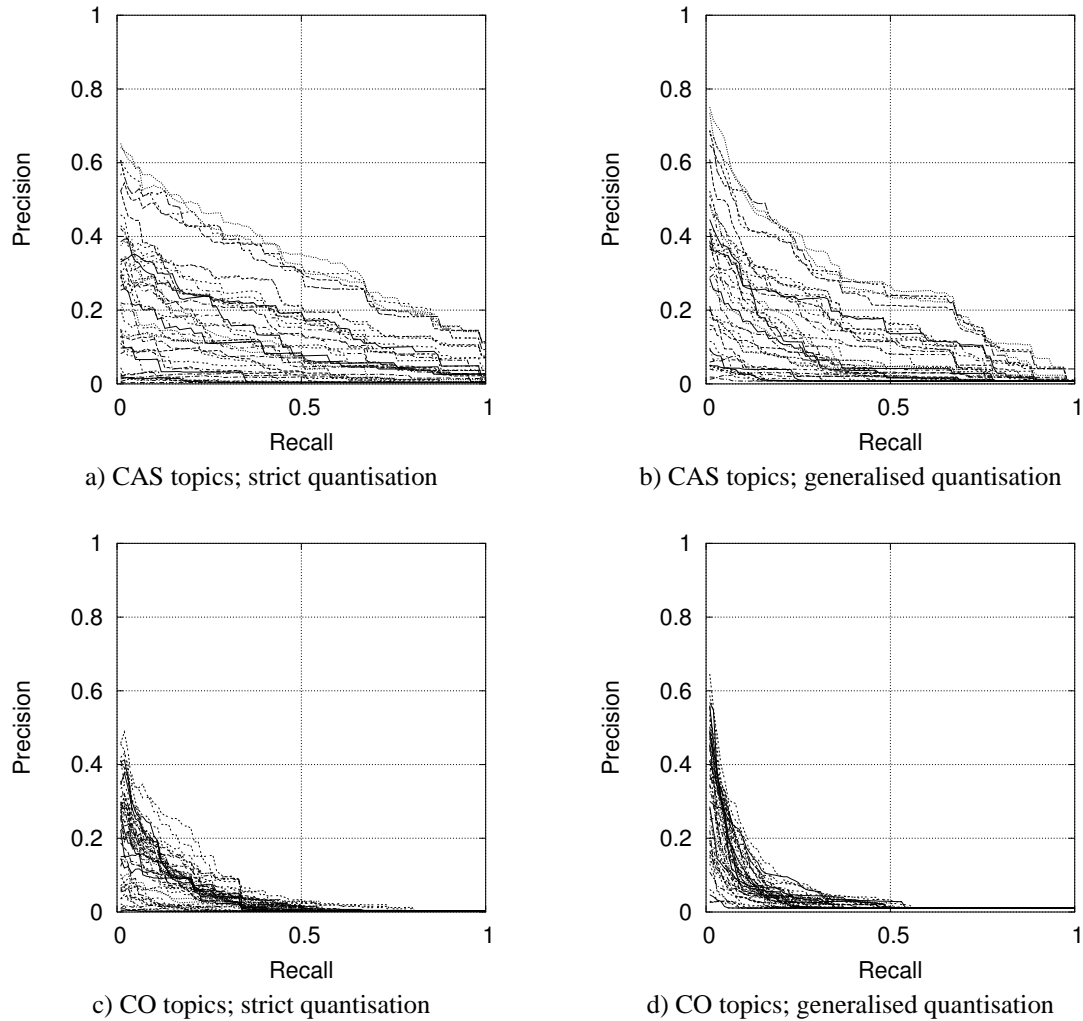


Figure 10: Summary of recall/precision curves for all INEX 2002 submissions

6 Summary of participants' results

For INEX 2002, a total of 51 runs (42 of them contained results for the CAS topics, 49 of them contained results for the CO topics) were submitted by 25 participating organisations. Figure 10 summarises the recall/precision graphs for CAS and CO topics, using the two quantisation functions f_{strict} and $f_{generalised}$.²

In addition to the recall/precision curves, the `inex_eval` software computes the average precision for 100 recall points. The submissions have been ranked according to the average precision. The top ten submissions for each task and each quantisation function are displayed in Table 7. Detailed evaluation results for the runs submitted for INEX 2002 can be obtained from [4].

When comparing the rankings for the two different quantisation functions it becomes evident that they are quite similar. A regression analysis based on average precision values for the submissions shows a strong linear correlation between results obtained using strict quantisation and results obtained using generalised quantisation. Figure 11 depicts the scatter plots for CAS and CO topics and the respective regression lines. For CAS topics the correlation coefficient is 0.9943, for CO topics 0.8875.

²All evaluation results have been compiled using the assessment package version 1.8 and `inex_eval` version 0.007.

rank	avg precision	organisation	run ID
1.	0.3438	CSIRO Mathematical and Information Sciences	manual
2.	0.3411	IBM Haifa Labs	Merge
3.	0.3248	IBM Haifa Labs	ManualNoMerge
4.	0.3093	IBM Haifa Labs	NoMerge
5.	0.3090	University of Michigan	no-duplicate
6.	0.3090	University of Michigan	allow-duplicate
7.	0.2257	University of Amsterdam	UAmsI02NGiSt
8.	0.2233	University of Amsterdam	UAmsI02NGram
9.	0.1865	University of California, Berkeley	Berkeley03
10.	0.1839	University of Amsterdam	UAmsI02Stem

a) CAS topics; strict quantisation

rank	avg precision	organisation	run ID
1.	0.2752	CSIRO Mathematical and Information Sciences	manual
2.	0.2706	IBM Haifa Labs	Merge
3.	0.2634	University of Michigan	allow-duplicate
4.	0.2634	University of Michigan	no-duplicate
5.	0.2535	IBM Haifa Labs	ManualNoMerge
6.	0.2419	IBM Haifa Labs	NoMerge
7.	0.1782	University of Amsterdam	UAmsI02NGiSt
8.	0.1770	University of Amsterdam	UAmsI02NGram
9.	0.1592	University of Amsterdam	UAmsI02Stem
10.	0.1583	Tarragon Consulting Corporation	tgnCAS_base

b) CAS topics; generalised quantisation

rank	avg precision	organisation	run ID
1.	0.0883	Universität Dortmund / Universität Duisburg-Essen	Epros03
2.	0.0809	Royal School of Library and Information Science	bag-of-words
3.	0.0670	Universität Dortmund / Universität Duisburg-Essen	Epros06
4.	0.0627	Queensland University of Technology	inexresult2.xml
5.	0.0592	University of Amsterdam	UAmsI02NGram
6.	0.0590	Queensland University of Technology	inexresults3.xml
7.	0.0556	Universität Dortmund / Universität Duisburg-Essen	plain hyrex
8.	0.0532	University of Amsterdam	UAmsI02NGiSt
9.	0.0520	Centrum voor Wiskunde en Informatica (CWI)	R_article
10.	0.0503	University of Minnesota Duluth	01

c) CO topics; strict quantisation

rank	avg precision	organisation	run ID
1.	0.0705	Universität Dortmund / Universität Duisburg-Essen	Epros03
2.	0.0635	Universität Dortmund / Universität Duisburg-Essen	Epros06
3.	0.0618	Royal School of Library and Information Science	bag-of-words
4.	0.0582	Sejong Cyber University	TitleKeywordsWLErr
5.	0.0572	Universität Dortmund / Universität Duisburg-Essen	plain hyrex
6.	0.0555	Centrum voor Wiskunde en Informatica (CWI)	R_article
7.	0.0554	University of Amsterdam	UAmsI02NGiSt
8.	0.0546	University of Amsterdam	UAmsI02NGram
9.	0.0499	University of Twente	utwente1pr
10.	0.0483	University of Melbourne	um_mgx2_long

d) CO topics; generalised quantisation

Table 7: Ranking of submissions w. r. t. average precision

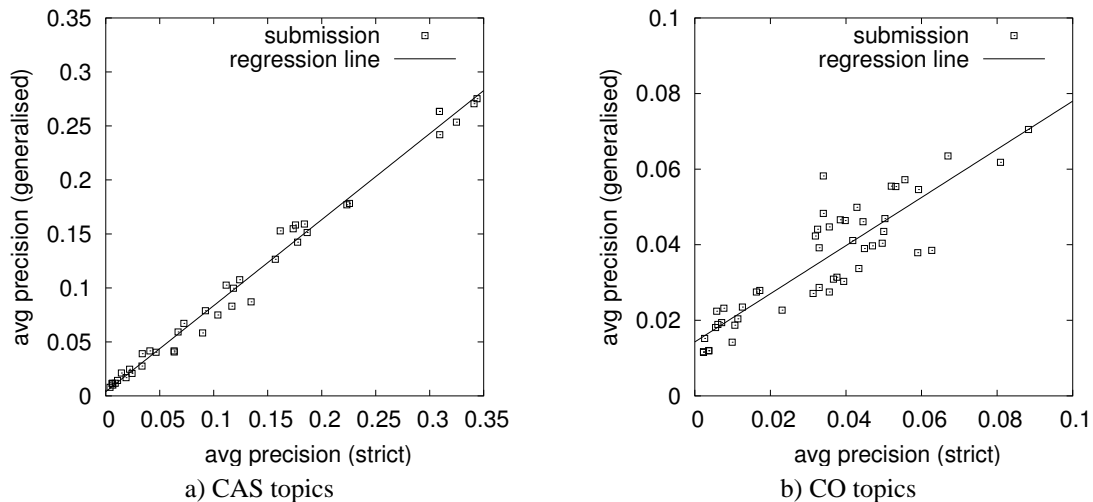


Figure 11: Scatter plots and regression lines for average precision of submissions, using strict and generalised quantisation.

7 Conclusions and outlook on INEX 2003

Within the first round of INEX in 2002, as a result of a collaborative effort with research groups from 36 different organisations worldwide, an infrastructure has been created for evaluating the effectiveness of content-oriented retrieval of XML documents. A document collection with real world XML documents from the IEEE Computer Society's digital library has been set up; 60 topics were created; the INEX 2002 participants provided assessments for 55 of these topics. Based on the notion of recall and precision, a metric for evaluating the effectiveness of XML retrieval has been developed and applied for evaluating the participants' submissions.

At the time of this writing, the call for participation in the INEX 2003 round has been published already. In 2003 we aim to extend the test collection with additional topics. The retrieval task, ad-hoc retrieval with CAS and CO topics, will remain the same. However, participants now can benefit from the test collection created in 2002 and optimise their retrieval approaches accordingly. We are looking forward to many participating organisations again with a broad range of retrieval approaches, thus promoting research in the field of XML retrieval.

8 Acknowledgements

We would like to thank the DELOS Network of Excellence for Digital Libraries³ for partially funding the INEX initiative. Special thanks go to the IEEE Computer Society⁴: Without their XML document collection INEX would not have happened. Additional acknowledgements go to *Deutscher Akademischer Austausch Dienst (DAAD)*⁵ and *The British Council*⁶ who supported INEX through their Academic Research Collaboration (ARC) Programme. Last but not least, we would like to thank the participating organisations and people for their contributions to the INEX test collection.

References

- [1] W. S. Cooper. Expected search length: A single measure of retrieval effectiveness based on weak ordering action of retrieval systems. *Journal of the American Society for Information Science*, 19:30–41, 1968.

³<http://delos-noe.org/>

⁴<http://computer.org/>

⁵<http://www.daad.de/>

⁶<http://www.britishcouncil.org/>

- [2] Norbert Fuhr, Norbert Gövert, Gabriella Kazai, and Mounia Lalmas. INEX: Initiative for the Evaluation of XML retrieval. In Ricardo Baeza-Yates, Norbert Fuhr, and Yoelle S. Maarek, editors, *Proceedings of the SIGIR 2002 Workshop on XML and Information Retrieval*, 2002.
- [3] Norbert Fuhr, Norbert Gövert, Gabriella Kazai, and Mounia 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, Sophia Antipolis, France, March 2003. ERCIM.
- [4] INEX. INEX 2002 evaluation results in detail. In Fuhr et al. [3].
- [5] INEX. INEX guidelines for topic development. In Fuhr et al. [3].
- [6] INEX. INEX relevance assessment guide. In Fuhr et al. [3].
- [7] INEX. INEX retrieval result submission format. In Fuhr et al. [3].
- [8] Jaana Kekäläinen and Kalvero Järvelin. Using graded relevance assessments in IR evaluation. *Journal of the American Society for Information Science and Technology*, 53(13), September 2002.
- [9] V. V. Raghavan, P. Bollmann, and G. S. Jung. A critical investigation of recall and precision as measures of retrieval system performance. *ACM Transactions on Information Systems*, 7(3):205–229, 1989.
- [10] Thomas Schütz. Retrieval of complex objects, considering SGML documents as example (in German). Master’s thesis, University of Dortmund, Computer Science Department, 1998.
- [11] K. Sparck Jones and C. J. van Rijsbergen. Report on the need for and provision of an “ideal” information retrieval test collection. Technical report, British Library Research and Development Report 5266, Computer Laboratory, University of Cambridge, 1975.
- [12] trec_eval. Evaluation techniques and measures. In Voorhees and Harman [13].
- [13] E. M. Voorhees and D. K. Harman, editors. *The Tenth Text REtrieval Conference (TREC 2001)*, Gaithersburg, MD, USA, 2002. NIST.