The INEX Evaluation Initiative

Gabriella Kazai[1], Norbert Gövert[2], Mounia Lalmas[3], and Norbert Fuhr[4]

[1] Queen Mary University of London gabs@dcs.qmul.ac.uk
[2] University of Dortmund goevert@ls6.cs.uni-dortmund.de
[3] Queen Mary University of London mounia@dcs.qmul.ac.uk
[4] University of Duisburg-Essen fuhr@uni-duisburg.de

1 Introduction

The widespread use of the extensible Markup Language (XML) on the Web and in Digital Libraries brought about an explosion in the development of XML tools, including systems to store and access XML content. As the number of these systems increases, so is the need to assess their benefit to users. The benefit to a given user depends largely on which aspects of the user's interaction with the system are being considered. These aspects, among others, include response time, required user effort, usability, and the system's ability to present the user with the desired information. Users then base their decision whether they are more satisfied with one system or another on a prioritised combination of these factors.

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 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. As a result of a collaborative effort, INEX created an XML test collection consisting of publications of the IEEE Computer Society, 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 [4].

In this chapter we provide an overview of the INEX evaluation initiative. Before we talk about INEX, we first take a brief look, in section 2, at the evaluation practices of information retrieval (IR) as these formed the basis of our work in INEX. In our discussion of INEX we follow the requirements that evaluations in IR are founded upon [10]. These include the specification of the evaluation objective (e.g. what to evaluate) in section 3, and the selection of suitable evaluation criteria in section 4. This is followed by an overview of the methodology for constructing the test collection in section 5. We describe the evaluation metrics in section 6. Finally we close with thoughts for future work in section 7.

2 Approaches to Evaluation

Evaluation means assessing the value of a system or product. It plays an important part in the development of retrieval systems as it stimulates their improvement. We can distinguish between comparative or goal-based evaluation approaches depending on whether a system is compared against others or is evaluated with respect to a given objective. Furthermore, evaluations may consider the individual components of a system or assess the system as a whole.

The evaluation of IR systems usually follows the comparative approach and considers a system in its entirety. A wealth of evaluation studies and initiatives to IR exist today. They can be classified into system and user-centred evaluations, and these further divided into engineering (e.g. efficiency), input (e.g. coverage), processing (e.g. effectiveness), output (e.g. presentation), user (e.g. user effort) and social (e.g. impact) levels [10, 2]. Most work in IR evaluation has been on system-centred evaluations and, in particular, at the processing level. At this level, the aspects most commonly under investigation are retrieval efficiency (e.g. speed, required storage) and retrieval effectiveness, i.e. the system's ability to satisfy the user's information need. For document retrieval systems, this is usually translated to the more specific criterion of a system's ability to retrieve in response to a user request as many relevant documents and as few non-relevant documents as possible.

The predominant approach to evaluate a system's retrieval effectiveness is with the use of test collections constructed specifically for that purpose. A test collection usually consists of a set of documents, user requests, and relevance assessments (i.e. the set of "right answers" for the user requests). There have been several large-scale evaluation projects, which resulted in established IR test collections [2, 9, 11]. One of the largest evaluation initiatives is the Text REtrieval Conference (TREC)[5], which every year since 1992, runs numerous tracks based on an increasingly diverse set of tasks that are to be performed on continually growing test collections [5, 12].

Besides a test collection, the evaluation of retrieval effectiveness also requires appropriate measures and metrics. A number of measures have been proposed over the years. The most commonly used are recall and precision. Recall corresponds to, in the above specification of effectiveness, to "a system's ability to retrieve as many relevant documents as possible", whereas precision pertains to "a system's ability retrieve as few non-relevant documents as possible". Given *Retr* as the set of retrieved documents and *Rel* as

[5] `http://trec.nist.gov/`

the set of relevant documents in the collection, recall and precision are defined as:

$$\text{recall} = \frac{|Rel \cap Retr|}{|Rel|} \qquad \text{precision} = \frac{|Rel \cap Retr|}{|Retr|} \tag{1}$$

Several metrics have been developed in order to apply these set-based measures to (possibly weakly ordered) rankings of documents. A recall/precision graph is typically used as a combined evaluation measure for retrieval systems. Such a graph, given an arbitrary recall value, plots the corresponding precision value. Raghavan et al.'s method [8] is based on the interpretation of precision as the probability $P(Rel|Retr)$ that a document viewed by a user is relevant. Assuming that the user stops viewing the ranking after a given number of relevant documents NR, precision is given as:

$$P(Rel|Retr)(NR) := \frac{NR}{NR + esl_{NR}} = \frac{NR}{NR + j + s \cdot i/(r+1)} \tag{2}$$

The expected search length, esl_{NR}, denotes the total number of non-relevant documents that are estimated to be retrieved until the NRth relevant documents is retrieved [3]. Let l denote the rank from which the NRth relevant document is drawn. Then j is the number of non-relevant documents within the ranks before rank l, s is the number of relevant documents to be taken from rank l, and r and i are the numbers of relevant and non-relevant documents in rank l, respectively.

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 documents with regard to the user request in the collection; $x \in [0, 1]$ denotes an arbitrary recall value):

$$P(Rel|Retr)(x) := \frac{x \cdot n}{x \cdot n + esl_{x \cdot n}} = \frac{x \cdot n}{x \cdot n + j + s \cdot i/(r+1)} \tag{3}$$

Based on this probabilistic interpretation of precision, recall/precision graphs can be established. Moreover, given that a system is to be evaluated with regard to multiple user requests, average precision can be calculated for a set of arbitrary recall points. Thus, a recall/precision graph defined for a set of multiple user requests can be defined.

Although IR research offers a wealth of evaluation measures, metrics and test collections, their application to the evaluation of content-oriented XML retrieval is limited due to the additional requirements introduced when the structure of XML documents is taken into account. This is illustrated by the retrieval paradigm implemented by XML retrieval systems, which, given a typical IR style information need, allows document components of varying granularity – instead of whole documents – to be returned to the user. Furthermore, users of XML retrieval systems are also able to issue queries that exploit the structure of the data and restrict the search to specific structural elements within an XML collection. Traditional IR test collections are not suitable for evaluating the retrieval effectiveness of content-oriented XML retrieval as they base their evaluation on the following implicit assumptions about the documents and the behaviour of users:

1. The relevance of one document is assumed to be independent of the relevance of any other documents in the collection.
2. A document is regarded as a well-distinguishable (separate) unit representing a retrievable entity.
3. Documents are considered as units of (approximately) equal size.
4. Given a ranked output list, the supposed user behaviour is that users look at one document after another and then stop at an arbitrary point. Thus, non-linear forms of output (e.g. sub-lists in Google) are not considered.

For content-oriented XML document retrieval, most of these assumptions do not hold and have to be revised:

1. Since arbitrary components of a document can be retrieved, multiple components from the same document can hardly be viewed as independent units.
2. XML documents consist of nested structures where document components of varying granularity may be retrieved, which cannot always be regarded as separate units.
3. The size of the retrieved components cannot be considered even approximately equal, but may vary from elements such as author names or paragraphs to complete documents or books.
4. When multiple components from the same document are retrieved, a linear ordering of the result items may confuse the user as these components may be interspersed with components from other documents. To address this issue some systems cluster components from the same document together, resulting in non-linear outputs.

The above assumptions also bear influence on the applied measures and metrics. For example, when computing precision at certain ranks, it is implicitly assumed that a user spends a constant time per document. Based on the implicit definition of effectiveness as the ratio of output quality vs. user effort, quality is measured for a fixed amount of effort in this case. However, the appropriateness of such a measure becomes questionable when the retrieved components significantly vary in size. It is therefore necessary to develop new measures and procedures for the evaluation of content-oriented XML retrieval. These and related issues were addressed in INEX and are further examined in the next sections.

3 Evaluation Objective

We set the evaluation objective as the assessment of a system's retrieval effectiveness, where we defined effectiveness as a measure of a system's ability to satisfy both content and structural requirements of a user's information need. Based on a document-centric view of XML, the above definition corresponds to the task of retrieving the most specific relevant document components, which are exhaustive to the topic of request [1].

The combination of content and structural aspects were also reflected in the task that was set to be performed by the participating groups: the ad-hoc retrieval of XML documents. The ad-hoc task is 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 [5]. Although XML retrieval systems can be applied to a number of tasks such as routing, filtering and interactive retrieval, this being the first year of the initiative, we decided to run only one track focusing on the ad-hoc task.

For the evaluation of the ad-hoc retrieval of XML documents, we needed to consider additional requirements brought upon by the extensive development and use of XML query languages. These query languages allow users to issue (directly or indirectly) complex queries that contain structural conditions. Consequently, the ability to service such queries must also be assessed by the evaluation process. On the other hand, content-oriented XML retrieval systems should also support queries that do not specify structural conditions as users are often not familiar with the exact structure of the XML documents. Taking this into account, we identified the following two types of queries:

Content-and-structure (CAS) queries are requests 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 elements of various complexity, e.g. at different levels of the XML documents' hierarchy.

Based on these two types of queries, we essentially defined two sub-tasks within the ad-hoc retrieval of XML documents. According to the latter subtask (using CO queries), effectiveness is measured as a system's ability to retrieve the most specific relevant document components, which are exhaustive to the topic of request. However, according to the sub-task based on CAS queries, a system's effectiveness is measured by its ability to retrieve the most specific relevant document components, which are exhaustive to the topic of request and match its structural constraints.

4 Evaluation Criteria

Traditional IR experiments designate relevance as a criterion for evaluating retrieval effectiveness. In INEX, retrieval effectiveness measures a combination of content and structural requirements. Relevance therefore is no longer sufficient as a single evaluation criterion, but has to be complemented with another dimension in order to allow reasoning about the structure. We chose the following two criteria:

Topical relevance, which is primarily a content related criterion. It reflects the extent to which the information contained in a document component satisfies the user's information need, e.g. measures the exhaustivity of the topic within a component.

Component coverage, which is a criterion that considers the structural aspects and reflects the extent to which a document component is focused on the information need, e.g. measures the specificity of a component with regards to the topic.

The basic threshold for relevance was defined as a piece of text that mentions the topic of request. A consequence of this definition is that container components of relevant document components in a nested XML structure are also regarded as relevant, albeit too large components. This clearly shows that relevance as a single criterion is not sufficient for the evaluation of content-oriented XML retrieval. Hence, the second dimension, component coverage, is used to provide a measure with respect to the size of a component by reflecting the ratio of relevant and irrelevant content within a document component. In actual fact, both dimensions are related to component size. For example, the more exhaustively a topic is discussed the more likely that the component is longer in length, and the more focused a component the more likely that it is smaller in size.

When considering the use of the above two criteria for the evaluation of XML retrieval systems, we must also decide about the scales of measurements to be used. For relevance, binary or multiple degree scales are known. In INEX, we chose a multiple degree relevance scale as it allows the explicit representation of how exhaustively a topic is discussed within a component with respect to its sub-components. For example, a section containing two paragraphs may be regarded more relevant than either of its paragraphs by themselves. Binary values of relevance cannot reflect this difference. We adopted the following four-point relevance scale [7]:

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.

For component coverage we used the following four-category nominal scale:

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 when retrieved by itself.

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 when retrieved by itself.

According to the above definition of coverage it becomes possible to reward XML engines that are able to retrieve the appropriate ("exact") sized document components. For example, a retrieval system that is able to locate the only relevant section in an encyclopaedia is likely to trigger higher user satisfaction than one that returns a too large component, such as the whole encyclopaedia. On the other hand, the above definition also allows the classification of components as too small if they do not bear self-explaining information for the user and thus cannot serve as informative units [1]. Take as an example, a small text fragment, such as the sentence "These results clearly show the advantages of content-oriented XML retrieval systems.", which, although part of a relevant section in a scientific report, is of no use to a user when retrieved without its context.

Only the combination of these two criteria allows the evaluation of systems that are able to retrieve components with high relevance and exact coverage, e.g. components that are exhaustive to and highly focused on the topic of request and hence represent the most appropriate components to be returned to the user.

5 Methodology for Constructing the Test Collection

The aim of a test collection construction methodology is to derive a set of queries and relevance assessments for a given document collection. The methodology for constructing a test collection for XML retrieval, although similar to that used for building traditional IR test collections, has additional requirements [6]. The following sections detail the processes involved and describe the resulting test collection.

5.1 Documents

The document collection consists of the fulltexts of 12 107 articles from 12 magazines and 6 transactions of the IEEE Computer Society's publications, covering the period of 1995–2002, and totalling 494 megabytes in size.

Although the collection is relatively small compared with TREC, it has a suitably complex XML structure (192 different content models in DTD) and contains scientific articles of varying length. On average, an article contains 1 532 XML nodes, where the average depth of a node is 6.9.

All documents of the collection are tagged using XML conforming to one common DTD. The overall structure of a typical article, shown in Figure 1, consists of a *front matter* (`<fm>`), a *body* (`<bdy>`), 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, and layout (such as emphasised and bold faced text), etc. The back matter contains a bibliography and further information about the article's authors.

```
<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>
    <sec>
      <st>...</st>
      ...
      <ss1>...</ss1>
      <ss1>...</ss1>
      ...
    </sec>
    ...
  </bdy>
  <bm>
    <bib>
      <bb>
        <au>...</au>
        <ti>...</ti>
        ...
      </bb>
      ...
    </bib>
  </bm>
</article>
```

Fig. 1. Sketch of the structure of the typical INEX articles

5.2 Participating Organisations

36 organisations from 15 countries on four continents participated in INEX 2002. Due to the diversity in the background of the participating groups a wide range of different approaches to XML retrieval were represented. We tried to classify these using the following three categories:

IR model-oriented: Research groups that focus on the extension of a specific type of IR model (e.g. vector space, rule-based, logistic regression), which they have applied to standard IR test collections in the past, to deal with XML documents. 15 groups belonged to this category.

Database-oriented: Groups that are working on extending database management systems to deal with semistructured data; most of these groups also incorporate uncertainty weights, thus producing ranked results. 2 groups followed this approach.

XML-specific: Groups that, instead of aiming to extend existing approaches towards XML, 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). 3 groups were classified under this category.

Whereas most of the retrieval approaches, a total of 20, were pure IR, database (DB), or XML-oriented, some groups combined elements from two categories: 3 groups developed IR+XML approaches, 2 groups followed DB+XML methods and 2 groups combined IR and DB models. The remaining 9 groups were classified as following other approaches.

5.3 Topics

The topics of the test collection were created by the participating groups. We asked each organisation to create sets of content-only (CO), and content-and-structure (CAS) 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 four-stage task [4].

During the first stage participants created an initial description of their information need without regard to system capabilities or collection peculiarities. During the collection exploration stage, using their own XML retrieval engines, participants evaluated their candidate topics against the document collection. Based on the retrieval results they then estimated the number of relevant components to the candidate topics. Finally, in the topic refinement stage the components of a topic were finalised ensuring coherency and that each component could be used in the experiments in a stand-alone fashion (e.g. retrieval using only the topic title or description).

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 (30 CAS and 30 CO) topics were selected into the final set. The selection was based on the combination of different criteria, like 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, eliminating topics that were considered too ambiguous or too difficult to judge, and selecting topics with at least 2, but no more than 20 relevant items in the top 25 retrieved components.

Figures 2 and 3 show examples for both types of topics. As it can be seen, the four main parts of an INEX topic are the topic title, topic description, narrative and keywords. The topic title serves as a summary of both content and structure related requirements of a user's information need. Apart from consisting of a number of keywords that best describe what the user is looking for, it allows the definition of containment conditions and target elements. Using containment conditions users can query with respect to the subject areas of specific components, for example, they can request that the abstract section of an article should highlight the "advantages of content-oriented XML retrieval". Target elements allow the specification of components that should be returned to the user.

According to the above requirements, a topic title 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 not. The keywords component of a topic was added as a means to keep a record of the list of search terms used for retrieval during topic development.

Table 1 shows some statistics on the final set of INEX topics. We classified the target and context elements of the final 30 CAS topics based on their content type, e.g. components that contain 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, or a mixture of fact and content containment conditions, e.g. specifying the publication year and/or the subject of the title, or specifying the author and the subject of some document components.

5.4 Assessments

The final set of topics were distributed back to the participating groups, who then used these topics to search the document collection. The actual queries

```
<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/atl</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>
```

Fig. 2. 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>
```

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

	CAS	CO
no of topics	30	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
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 fact and content <ce>	12	–
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	–
avg no of words in topic description	18.8	16.1
avg no of words in keywords component	7.06	8.7

Table 1. Statistics on CAS and CO topics in the INEX test collection

put to the search engines had to be automatically generated from any part of the topics except the narrative. As a result of the retrieval sessions, groups produced ranked lists of XML elements in answer to each query. 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. A result element in a retrieval run was identified using a combination of file names and XPaths. The file name and file path uniquely identified an article within the document collection, and XPath allowed the location of a given node within the XML tree of the article. Associated with a result element were its retrieval rank and/or its relevance status value [4].

A total of 51 runs were submitted from 25 groups. For each topic, the results from the submissions were merged to form the pool for assessment [12]. The assessment pools contained between one to two thousand document components from 300–900 articles, depending on the topic. The result elements varied from author, title and paragraph elements through sub-section and section elements to complete articles and even journals. The assessment pools were then assigned to groups 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 assessments were done along the two dimensions of topical relevance and component coverage. Assessments were recorded using an on-line assessment system, which allowed users to view the pooled result set of a given topic listed in alphabetical order, 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 [4].

Table 2 shows a summary of the collected assessments for CAS and CO topics. Note that these figures are based on the assessments of 53 of the 60 topics, as no groups volunteered to assess 5 of the topics and the assessment of a further 2 topics were not yet complete at the time of writing. The table shows a relatively large proportion of non-article level elements with exact coverage compared with article elements, which indicates that for most topics sub-components were considered as the preferred units to be returned to the user.

Rel+	Articles		Non-articles	
Cov	CAS	CO	CAS	CO
3E	187	309	2257	1087
2E	59	165	1128	1107
1E	82	114	1770	827
3L	173	394	424	1145
2L	137	599	507	2295
1L	236	854	719	2708
2S	21	118	846	3825
1S	54	116	1119	3156
0N	9430	9671	14860	7917
All	10379	12340	23630	24067

Table 2. Assessments at article and component levels

6 Evaluation Metrics

Due to the nature of XML retrieval, it was necessary to develop new evaluation procedures. These were based on the traditional recall/precision and, in particular, the metrics described in section 2. However, before we could apply these measures, we first had to derive a single relevance value based on the two dimensions of topical relevance and component coverage. For this purpose we defined a number of quantisation functions, $\mathbf{f}_{quant}$:

$$\begin{aligned} \mathbf{f}_{quant} : \mathit{Relevance} \times \mathit{Coverage} &\rightarrow [0,1] \\ (rel, cov) &\mapsto \mathbf{f}_{quant}(rel, cov) \end{aligned} \tag{4}$$

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

The rational behind such a quantisation function is that overall relevance of a document component can only be determined using the combination of relevance and coverage assessments. Quantisation functions can be selected according to the desired user standpoint. For INEX 2002, two different functions have been selected: $\mathbf{f}_{strict}$ and $\mathbf{f}_{generalised}$. The quantisation function $\mathbf{f}_{strict}$ is used to evaluate whether a given retrieval method is capable of retrieving highly relevant and highly focused document components:

$$\mathbf{f}_{strict}(rel, cov) := \begin{cases} 1 & \text{if } rel = 3 \text{ and } cov = \text{E}, \\ 0 & \text{else} \end{cases} \tag{5}$$

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

$$\mathbf{f}_{generalised}(rel, cov) := \begin{cases} 1.00 & \text{if } (rel, cov) = \text{3E}, \\ 0.75 & \text{if } (rel, cov) \in \{\text{2E}, \text{3L}\}, \\ 0.50 & \text{if } (rel, cov) \in \{\text{1E}, \text{2L}, \text{2S}\}, \\ 0.25 & \text{if } (rel, cov) \in \{\text{1S}, \text{1L}\}, \\ 0.00 & \text{if } (rel, cov) = \text{0N} \end{cases} \tag{6}$$

Given this type of quantisation, 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 Raghavan et al.'s evaluation procedure could be applied directly. To apply equation 3 for $\mathbf{f}_{generalised}$ the variables n, j, i, r, and s are to be interpreted as expectations. For example, given a function $assessment(c)$, which yields the relevance/coverage assessment for a given document component c, the number n of relevant components with respect to a given topic is computed as:

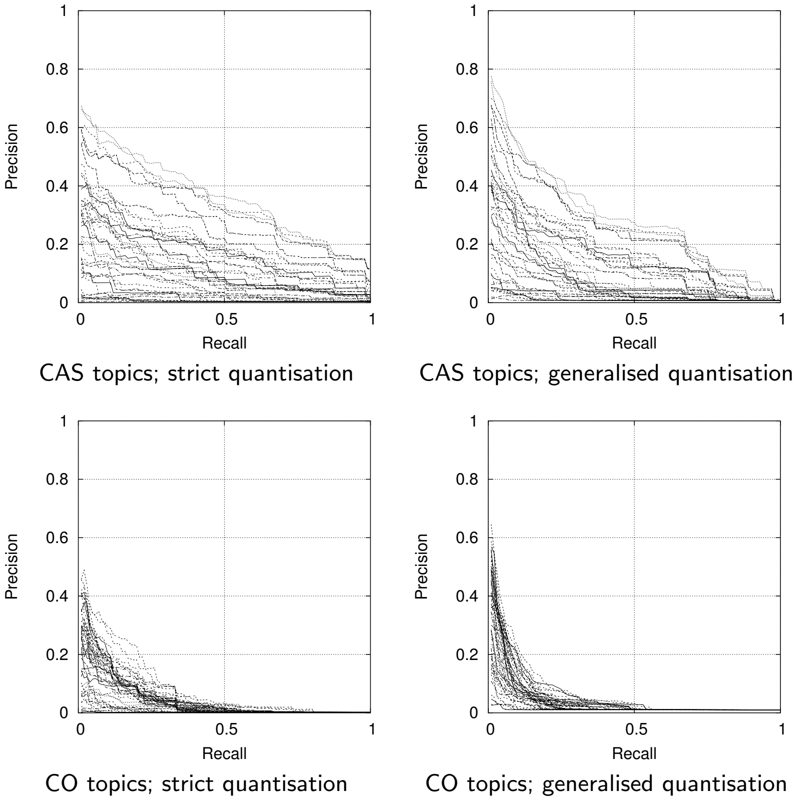
$$n = \sum_{c \in components} \mathbf{f}_{generalised}(\text{assessment}(c)) \tag{7}$$

Expectations for the other variables are computed respectively.

Figure 4 displays the recall/precision graphs obtained from the 51 submissions of 25 groups, using the strict and generalised quantisation functions (see [4] for the detailed evaluation results).

7 Conclusions and Future Work

As a collaborative effort of research groups from 36 organisations worldwide, the INEX evaluation initiative in 2002 created an infrastructure for evaluating the effectiveness of content-oriented retrieval of XML documents. A document



CAS topics; strict quantisation

CAS topics; generalised quantisation

CO topics; strict quantisation

CO topics; generalised quantisation

Fig. 4. Summary of recall/precision curves for all INEX 2002 submissions

collection with real life XML documents from the IEEE Computer Society's digital library has been set up, 60 topics created and assessments provided for 53 of these topics. Based on the notion of recall and precision, metrics for evaluating the effectiveness of XML retrieval have also been developed. These were applied to evaluate the submitted retrieval runs of the participating groups.

In the second round of INEX, commencing from April 2003, we aim to extend the test collection and develop alternative evaluation measures and metrics addressing the issue of overlapping result elements. We are also working on a new topic format, which will allow the representation of vague structural conditions. In the long term future of INEX we aim to extend the range of tasks under investigation to include, in particular, interactive retrieval, which will be based on new evaluation criteria reflecting typical user interaction with structured documents.

Acknowledgements

We would like to thank the DELOS Network of Excellence for Digital Libraries[6] for partially funding the INEX initiative. Special thanks go to the IEEE Computer Society[7]. Without their XML document collection INEX would not have happened. Additional acknowledgements go to *Deutscher Akadmischer Austausch Dienst (DAAD)*[8] and *The British Council*[9] 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] Y. Chiaramella, P. Mulhem, and F. Fourel. A Model for Multimedia Information Retrieval. Technical report, FERMI ESPRIT BRA 8134, University of Glasgow, April 1996.

[2] C.W. Cleverdon, J. Mills, and E.M. Keen. Factors Determining the Performance of Indexing Systems, Vol. 2: Test Results. Technical report, Aslib Cranfield Research Project, Cranfield, England, 1966.

[3] 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.

[4] 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.

[5] D. Harman. The TREC conferences. In R. Kuhlen and M. Rittberger, editors, *Hypertext - Information Retrieval - Multimedia, Synergieeffekte elektronischer Informationssysteme, Proceedings HIM '95*, volume 20 of *Schriften zur Informationswissenschaft*, pages 9–28, Konstanz, April 1995. Universitätsverlag Konstanz.

[6] G. Kazai, M. Lalmas, and J. Reid. Construction of a Test Collection for the Focussed Retrieval of Structured Documents. In *25th European Conferenve on Information Retrieval Research (ECIR 2003)*, March 2003.

[7] 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.

[6] `http://delos-noe.org/`

[7] `http://computer.org/`

[8] `http://www.daad.de/`

[9] `http://www.britishcouncil.org/`

[8] 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.

[9] G. Salton, editor. *The SMART Retrieval System - Experiments in Automatic Document Processing*. Prentice Hall, Englewood, Cliffs, New Jersey, 1971.

[10] Tefko Saracevic. Evaluation of evaluation in information retrieval. In E.A. Fox, P. Ingwersen, and R. Fidel, editors, *Proceedings of the 18th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 138–146, New York, 1995. ACM. ISBN 0-89791-714-6.

[11] W. M. Shaw, J. B. Wood, R. E. Wood, and H. R. Tibbo. The cystic fibrosis database: Content and research opportunities. *Library and Information Science Research*, 13:347–366, 1991.

[12] E. M. Voorhees and D. K. Harman, editors. *The Tenth Text REtrieval Conference (TREC-2001)*, Gaithersburg, MD, USA, 2002. NIST.