INEX: Initiative for the Evaluation of XML Retrieval

Norbert Fuhr Norbert Gövert
University of Dortmund, Germany

Gabriella Kazai Mounia Lalmas
Queen Mary University of London, UK

http://qmir.dcs.qmw.ac.uk/INEX/

The widespread use of XML prompted the development of appropriate searching and browsing methods for XML documents. This explosion of XML retrieval tools requires the development of appropriate testbeds and evaluation methods. As part of a large-scale effort to improve the efficiency of research in information retrieval and digital libraries, the INEX initiative organises an international, coordinated effort to promote evaluation procedures for content-based XML retrieval. The project provides an opportunity for participants to evaluate their retrieval methods using uniform scoring procedures and a forum for participating organisations to compare their results.

1 Introduction

XML is going to become the standard document format in digital libraries, product catalogues, scientific data repositories and across the Web. The major purpose of XML markup is the explicit representation of the logical structure of a document.

A closer look at the different types of applications as well as on the standards under development reveals that there are in fact two different views on XML:

- The *document-centric view* focuses on structured documents in the traditional sense, where XML is used for logical markup of texts both at the macro level (e. g. chapter, section, paragraph) and the micro level (e. g. MathML for mathematical formulas, CML for chemical formulas). XML DTDs, namespaces, XPath and XSL are W3C standards[1] based on this view.

- The *data-centric view* uses XML for exchanging formatted data in a generic, serialised form between different applications (e. g. spreadsheets, database records). This is especially important for the interoperability of Web services (e. g. e-business applications). XML schema and the proposed XML query language XQuery address the data-centric issues of XML.

[1] http://www.w3.org/

For testing systems based on the data-centric view, synthetic XML data (produced e. g. by IBM's XML generator[2]) can be used; also certain document-oriented applications may be evaluated this way.

However, for testing content-based access methods, only real world documents with real world uses are suitable, as it is the case in the major evaluation initiatives for flat document retrieval, like TREC[3] and CLEF[4]. The Initiative for Evaluation of XML Retrieval (INEX) aims at building such a testbed for XML documents.

During the FOCUS project[5], a variety of document collections were considered with respect to their suitability as a testbed for XML retrieval, but none of them fulfilled the criteria with respect to collection size and document structure. Finally, we came across a collection of journal articles (in SGML format) from IEEE Computer Society, which is sold as a CD set[6]. When we contacted IEEE-CS directly, they generously offered us seven years of their journal publications, which were already converted into XML format.

Besides a document collection, a test suite also must contain a set of uses (or tasks) along with users being able to judge about the quality of answers returned by the system. Due to limited funding for INEX, there was no money for paying subject specialists for making relevance judgements. Since the collection is from the area of computer science, we recruited our users from the participating research groups to formulate some queries, for which they later also have to provide appropriate relevance judgements.

Following the call for participation in March 2002, 49 research groups registered within six weeks. This number shows the large interest in this research area. As the time of this writing, participating groups have submitted their queries, from which the final set of queries has been selected and distributed back to the participants. Currently, the participants create their retrieval runs. Relevance judgements are performed during fall 2002. After collecting the judgements and computation of the quality of the different participating systems, the INEX initiative will conclude with a final workshop in December 2002.

In the remainder of this paper, we first describe the document collection (Section 2), and the type of tasks to be performed on this data (Section 3). Then we give a brief survey over the participating groups and their research interests (Section 4). Finally, we point out some issues that are to be solved while the initiative is running.

2 Document Collection

The documents of the INEX test collection are selected as being a sample of texts occurring in operational retrieval environments with content-based access to semi-structured documents. The fulltexts of IEEE Computer Society's publications[7] from 1995 up to the beginning of 2002 form the document collection of INEX. The collection includes the articles of the respective volumes of 12 magazines and 6 transactions. Table 1 shows some statistics of the document collection.

While the collection is relatively small in terms of the number of documents it has reasonable size if one considers the total volume of the collection of about 494 megabytes. The complexity of the collection can also be seen from the parameters describing the XML structure of the articles.

[2] `http://www.alphaworks.ibm.com/tech/xmlgenerator/`

[3] `http://trec.nist.org/`

[4] `http://www.clef-campaign.org/`

[5] `http://ls6-www.cs.uni-dortmund.de/ir/projects/focus/`, funded by the German DAAD and the British Council

[6] `http://www.computer.org/cspress/catalog/cs-96.htm`

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

no of articles	12 107
size in MB	494
average no of bytes / article	42 758
no of content models in DTD	192
avg no of nodes / article	1 532
avg path length	6.9

Table 1: Statistics of the INEX document collection

A single article contains an average of 1 532 XML element, attribute, and text nodes. The DTD contains 192 different content models. The average depth of an XML node with content (attribute and text nodes) is 6.9.

```
<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>
```

Figure 1: Sketch of the structure of the INEX documents

Figure 1 shows an excerpt of the structure of one of the documents of the collection. The overall structure of a typical article is as follows: it 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 is the article's body which contains the content, structured in sections (`<sec>`), sub sections (`<ss1>`), and sub sub section (`<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 face), etc. The back matter contains a bibliography and information about the article's authors.

3 Uses

With the use of XML query languages, users of XML retrieval systems are able to exploit the structural nature of the data and restrict their search to specific structural elements within an XML collection. The content-based retrieval of XML documents, 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 and stems from the fact that users often do not know the exact structure of the XML documents, or may not want to restrict their search to specific structural elements. Taking this into account we identified two types of queries to be included in the INEX test collection:

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 any elements of the XML collection that fullfill the query.

Examples of both types of topic are given in Figures 2 and 3.

The example of the CAS topic shows that the target elements (`<te>`), e. g. what the user wants to retrieve, are defined explicitly and, in this case, are whole article elements. Target elements can be either attribute-like (metadata) elements, such as author, title, publication information, or content-like elements, such as article, section, paragraph. Apart from the target elements, the example also specifies a number of concept-context pairs (`<cw>`, `<ce>`), which define concepts that context elements should be about. Similarly to target elements, we can distinguish two types of concept-context conditions: attribute-like and content-like. Attribute-like conditions, such as the publication year, have a stronger data-centric view on XML documents, whereas content-like conditions, such as a section of the article's body being about "non-monotonic reasoning", are closer to the document-centric view of XML. The example contains an 'unpaired' concept, "belief revision", which may be the subject of any XML element in the article.

The example of the CO topic shows that only the concepts, that documents or document components should be about, are specified. It places no restrictions as to what element types should be returned to the user or in which elements the search concepts should occur. Since no target element is defined, in XML retrieval it is the task of the search engine to identify the appropriate level of granularity that is to be returned to the user.

3.1 Topic Format

The INEX topic format is summarised in the topic DTD, shown in Figure 4. It follows the topic format developed for TREC, with the addition of the keywords component and the modification of the topic title to allow for the representation of structural conditions. An INEX topic therefore has four main parts: title, description, narrative and keywords, where the topic title is composed of one or more sub-parts.

A topic title is a short version of the topic statement, made up of concepts (listed within the `<cw>` component) that best describe what the user is looking for. In CAS queries, a topic title also specifies the target elements (`<te>`) of the search and / or the context elements (`<ce>`) of given search concepts (`<cw>`). Both target and context elements may list one or more XML elements

```
<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>
```

Figure 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>
```

Figure 3: A CO topic from the INEX test collection

(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 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.

A topic description is a one- or two-sentence natural language definition of an information need. The narrative is the explanation of the topic statement in more detail and the description of what makes a document / component relevant or not. Keywords are a record of the list of search terms used for retrieval during the topic development process (described in Section 3.2).

A topic also has three attributes: `topic-id` (1–60), `query-type` (CAS or CO) and `ct-no` (candidate topic number, 1–143).

```
<!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 4: Topic DTD for the INEX test collection

3.2 Topic Development

Within the INEX initiative it was the task of the participating organisations to provide the topics that now contribute the queries of the INEX test collection. We asked each organisation to create topics taking into account a number of factors to ensure the diversity of topics and that they are representative of what real users might ask and the type of the service that operational systems may provide. It was also a requirement that the author of a topic be familiar with some subject areas covered in the collection.

Participants were provided with guidelines identifying three stages of the topic creation process:

1. Creation of the initial topic statement, during which the initial information need is formed regardless of collection peculiarities to avoid artificial or collection-biased queries.

2. Collection exploration, during which the number of relevant XML documents / components to a topic is estimated. Unlike TREC, we didn't provide topic authors a retrieval system for this task, but participants performed retrieval on their candidate topics, using their own retrieval engine, and judged the top 25 and 100 retrieved components.

   Keywords used for the retrieval run were recorded and added to the keywords component of the topic.

3. Topic refinement, during which the components of a topic were finalised, to ensure coherency and that each component could be used in a stand-alone fashion (e. g. retrieval using only the topic title).

As a result of the topic creation task, a total of 143 candidate topics were submitted by the participants, of which 60 (30 CAS and 30 CO) were selected to be included in the INEX test collection. Due to the varied backgrounds of the participating organisations and their particular interests in XML retrieval, the submitted topics formed a diverse set of queries both with regards to their content and difficulty level. The selection of the final 60 topics was based on a number of criteria to ensure that we obtain a balanced set of test queries.

- Equal numbers of CO and CAS queries,
- Representative of both traditional and new search requirements for XML documents,
- Topics with reasonable information need description,
- Topics with possible multiple relevant element types,
- Topics with at least 2, but no more than 20 relevant items in the top 25 retrieved components.

	CAS	CO
no of queries	30	30
avg no of `<cw>` subparts / title	2.1	1
avg no of concepts / cw	2.5	4.3
avg of the total no of concepts / title	5.3	4.3
avg no of `<ce>` subparts / title	1.7	–
avg no of XML elements / `<ce>`	1.7	–
avg of the total no of XML elements / title	2.9	–
avg no of `<te>` subparts / title	1	–
avg no of XML elements / `<te>`	1.4	–
total no of attribute-like target elements	13	–
total no of content-like target elements	11	–
total no of topics with no target element specified	6	30
avg no of concept-context pairs / title	1.6	–
total no of attribute-like concept-context pairs	37	–
total no of content-like concept-context pairs	19	–
topics with only attribute concept-context conditions	5	–
topics with only content concept-context conditions	4	–
avg no of words in topic description	18.8	16.1

Table 2: Statistics on CAS and CO queries in the INEX test collection

Table 2 shows some statistics on the final set of queries. The final set of CAS queries contains a mixture of attribute-like and content-like target element and concept-context pair specifications. The target elements of 43 % of the CAS topics are attribute-like elements, such as author or title, while 37 % requests articles or article sub-elements to be returned, and 20 % have no target elements specified. Regarding the distribution of concept-context pairs, the majority (70 %) of the CAS queries contains both attribute-like and content-like concept-context conditions while the remaining 30 % contain only attribute-like or only content-like concept-context conditions. Two of the 60 queries contain negation.

3.3 Retrieval and Relevance Assessment

The task, to be performed with the data and the 60 topics, will be the ad-hoc retrieval of XML documents using automatic queries. The queries used in a retrieval session can be generated from any parts of the topics, but the narrative, using an automatic process. The answer to a query will be a ranked list of XML elements, the top 100 elements of which will be submitted as the retrieval result. Organisations may submit up to three retrieval runs. The retrieval results of each participating group will then be pooled and the top 1000 elements returned to the participants for relevance assessment.

The relevance assessments will be provided by the participating groups and should be made wherever possible by the topic authors.

4 Systems

Country	no of groups	Country	no of groups
USA	11	Finland	1
Canada	1	France	3
America	**12**	Germany	5
Australia	**3**	Greece	1
China	1	Ireland	2
India	1	Italy	3
Israel	1	Netherlands	4
Japan	1	Portugal	1
Korea	2	Spain	1
Asia	**6**	Switzerland	2
Austria	1	UK	3
Denmark	1	**Europe**	**28**

Table 3: Geographic distribution of the 49 registered research groups

Table 3 gives a survey over the geographic distribution of the participating groups. Given that this is the first (but hopefully not the last) year of INEX, it is quite surprising that we have participants from 21 countries on four continents. Also, European participation is stronger than in similar activities (e. g. TREC).

For registration, we asked the participants for a short description of their research approach towards XML retrieval. Although these descriptions are rather heterogeneous, we tried to classify them into four major categories:

IR model-oriented (20 groups): Most research groups focus on a specific type of IR model (e. g. vector space, rule-based, logistic regression, LSI) which they have applied to standard IR test collections in the past. Now, they are working on the extension of these models for dealing with XML documents.

Database-oriented (5 groups): In the database area, there is growing interest in extending database management systems such that they can deal with so-called semistructured data; for this purpose, most of the participating groups in this area also aim at incorporating uncertainty weights, thus producing ranked results.

XML-specific (14 groups): Whereas groups of the former two categories aim at extending existing approaches towards XML, a large number of research groups has developed models and systems specifically for XML. These groups have very different backgrounds; some of them start from XML standards (like XSL, XPath or XQuery), whereas other have developed new models specifically for XML.

System / data structure (9 groups): Finally, there are several groups who are more interested in system-oriented aspects, like e. g. developing new data structures and algorithms for XML, or by enhancing a standard IR system such that it supports also XML data.

5 Conclusions and Outlook

Evaluation of the retrieval effectiveness of the retrieval engines used by the participants will be based on the constructed test collection and uniform scoring techniques, including recall / precision measures, which take into account the structural nature of XML documents. Other measures, which consider "near misses", when an element near one that has been assessed relevant has been retrieved, will also be used. Discussion on appropriate evaluation metrics are ongoing.

The results of the evaluation will be returned to all participants. Participating organisations will present their approaches and compare their results at the workshop in December. All results will be published in the proceedings of the workshop and on the Web.

INEX is designed to be a long-term initiative with workshops held on a yearly basis, such that it becomes a valuable infrastructure for the evaluation of XML retrieval approaches.

Acknowledgements

The INEX initiative is sponsored by the DELOS Network of Excellence for Digital Libraries[8]. We also wish to thank the IEEE Computer Society[9] for providing us with the XML collection without any cost and Ellen Vorhees from NIST for providing the TREC guidelines for topic generation.

[8] http://ercim.org/DELOS/
[9] http://www.computer.org/