A report on the first year of the INitiative for the Evaluation of XML Retrieval (INEX'02)

Gabriella Kazai
Department of Computer Science, Queen Mary University of London. Email address: gabs@dcs.qmul.ac.uk

Mounia Lalmas
Department of Computer Science, Queen Mary University of London. Email address: mounia@dcs.qmul.ac.uk

Norbert Fuhr
Institute of Informatics and Interactive Systems, University of Duisburg-Essen. Email address: fuhr@uni-duisburg.de

Norbert Gövert
Department of Computer Science, University of Dortmund. Email address: goevert@ls6.cs.uni-dortmund.de

The INitiative for the Evaluation of XML retrieval (INEX) aims at providing an infrastructure to evaluate the effectiveness of content-oriented XML retrieval systems. To this aim, in the first round of INEX, in 2002, a test collection of real world XML documents along with a set of topics and respective relevance assessments has been created with the collaboration of 36 participating organisations. In this article, we provide an overview of the first round of the INEX initiative.

1. Introduction

With the widespread use of the eXtensible Markup Language (XML) on the Web and in Digital Libraries, XML is becoming the standard document representation format. In contrast to HTML, which is layout-oriented, XML follows the concept of separating a document's logical structure (using macro-level markup for chapters, sections, paragraphs, etc.) and semantics (based on micro-level markup, such as MathML for mathematical formulas, CML for chemical formulas, etc.) from its layout. From an information retrieval (IR) point of view, this enables users to benefit from the explicit structural and semantic information inherent in XML documents.

XML retrieval systems have been and are being developed to implement content-oriented retrieval approaches to XML documents, which support querying both with respect to content and structure (Blanken et al, 2003). These systems build on a combination of data retrieval techniques (as defined by the XPath and XQuery standards) and IR concepts in order to support the ranking of XML elements by their probability of relevance to a query (Baeza-Yates et al, 2000; Baeza-Yates et al, 2002; Fuhr et al, 2003). Another common feature of these approaches is that they follow a more focused retrieval paradigm from traditional IR. That is, instead of retrieving whole documents, systems aim at retrieving document components (e.g. XML elements of varying granularity) that fulfill the user's query.

As the number of XML retrieval systems increases, so is the need to assess their benefit to the user. The benefit is evaluated by considering the different aspects of the user's interaction with the system, such as usability, required user effort, response time, and the system's ability to present the user with the desired information. Within IR research, the evaluation of these aspects to a system's performance has a long and rich history that resulted in a wealth of evaluation studies and initiatives. These are usually classified into system- and user-centred evaluations, which are 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 (Saracevic, 1995; Cleverdon et al, 1966).

Most work in IR evaluation has been on system-centred evaluations and, in particular, at the processing 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 a user's query. For document retrieval systems, the latter is usually translated to the more specific criterion of a system's ability to retrieve in response to a user query as many relevant documents and as few non-relevant documents as possible. The predominant approach in IR to evaluate a system's retrieval effectiveness is with the use of test collections usually consisting of a set of documents, queries (topics), and relevance assessments (Voorhees and Harman, 2002).

Traditional IR test collections, however, are not suitable for the evaluation of content-oriented XML retrieval as they do not consider the structural information in XML collections, and base their evaluation on relevance assessments provided at the document level only. A closer look at the underlying principles of IR evaluation and the additional requirements introduced when the structure of XML documents is taken into account also reveals that XML retrieval does not comply with many of the implicit assumptions that IR evaluations are based

upon. For example, the assumption in IR that documents are independent units (whose relevance is independent of any other document) can no longer be treated as reasonable approximation in XML retrieval since multiple components retrieved from the same document can hardly be viewed as independent. Furthermore, when computing typical IR benchmarks, such as precision at certain ranks, it is implicitly assumed that users spend a constant time per document, where documents are assumed to be of approximately equal length. Since in XML retrieval, arbitrary document components of varying granularity may be retrieved (e.g. titles, paragraphs, sections or whole documents), the size of these components cannot be assumed to be even approximately equal. In addition, the supposed behaviour of users of IR systems is to look at documents one after the other, in a linear order, from the ranked output list. XML retrieval systems, however, may produce non-linear output lists in an effort to group related document components together (e.g. components from the same document) to minimize user disorientation. The invalidity of these assumptions in XML retrieval makes it necessary to build new test collections and develop new measures and procedures for the evaluation of XML retrieval systems.

To address these and related issues, the INEX evaluation initiative 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, with contributions from 36 participating organizations, INEX'02 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 (Fuhr et al, 2003).

In this paper we provide an overview of the first year of the INEX evaluation initiative[1]. The paper is structured as follows. In Section 2, we describe the evaluation objective and criteria that were adopted in INEX. In Section 3, we discuss the methodology used to construct the test collection. In Section 4, a specification of the evaluation metrics applied for INEX'02 is given. We end with conclusions and an outlook on INEX'03 in Section 5.

2. Setting up the initiative

In order to setup an evaluation initiative we must specify the objective of the evaluation (e.g. what to evaluate), select suitable criteria, set up measures, measuring instruments and a methodology (e.g. framework and procedures) (Saracevic, 1995). In traditional IR evaluations (processing level) the objective is to assess the retrieval effectiveness of IR systems, the criteria is relevance, the measures are recall and precision and the measuring instruments are relevance judgements. However, as it was pointed out in the previous section, these criteria and measures rely on implicit assumptions about the documents and users, which do not hold for XML retrieval. How these and related issues were addressed in INEX are described in the next two sub-sections.

2.1. Evaluation objective and task

As in traditional IR evaluations, INEX set as its objective the evaluation of a system's retrieval effectiveness, which, to take into account the structural aspects of XML retrieval, has been redefined as a measure of a system's ability to satisfy both content and structural requirements of a user's query. Based on the content-oriented 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.

Before we can evaluate a system's retrieval effectiveness, we also need to specify a retrieval task, which is to be performed by the participating groups. We set this task to be the ad-hoc retrieval of XML documents. Similarly to IR, we regard ad-hoc retrieval as a simulation of how a library might be used, where a static set of documents is searched using a new set of queries (topics) (Harman, 1995). The main differences are that, in INEX, the library consists of XML documents, the queries may contain both content and structural conditions and, in response to a query, arbitrary XML elements may be retrieved from the library.

An important point to emphasize here is the expressiveness of the developed XML query languages, which allow users of XML retrieval systems to issue complex queries that contain structural conditions. In order to consider the evaluation of the additional functionality introduced by these query languages, we defined two sub-tasks within the umbrella of the ad-hoc retrieval task:

1. The first sub-task centers around content-only (CO) queries, which are IR-style user requests that ignore the document structure. In this task, it is left to the retrieval system to identify the most appropriate XML elements to return to the user.
2. The second sub-task is based on content-and-structure (CAS) queries, which are requests that contain explicit references to the XML structure, either by restricting the context of interest or the context of certain search concepts.

With respect to the sub-task based on 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, with respect to the sub-task defined by CAS queries, a system's effectiveness is

[1] An extended version of the paper can be found in (Kazai et al, 2003).

measured by its ability to retrieve the most specific relevant document components, which are exhaustive to the topic of request and match the structural constraints specified in the query.

2.2. Evaluation criteria

Traditional IR experiments designate relevance as a criterion for evaluating retrieval effectiveness. In INEX, retrieval effectiveness is associated with the 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 document structure. We chose the following two criteria:

- Topical relevance, which reflects the extent to which the information contained in a document component satisfies the user's query, e.g. measures the exhaustivity of the topic within a component.
- Component coverage, which reflects the extent to which a document component is focused on the query, e.g. measures the specificity of a component with regards to the topic.

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. 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. We adopted the following four-point relevance scale (Kekäläinen and Järvelin, 2002):

- 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 definitions, the basic threshold for topical relevance is a piece of text that mentions the topic of request (Harman, 1995). A consequence of this definition is that container components of relevant document components in a nested XML structure, albeit too large components, are also regarded as relevant. However, our aim is to be able to differentiate retrieval systems that are able to locate, for example, the only relevant section in an encyclopedia from those that would return the whole encyclopedia, as the former is likely to trigger higher user satisfaction. 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. Component coverage 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. Based on the combination of the two criteria it becomes possible to reward 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 units to be returned to the user.

3. Constructing the Test Collection

The following sections describe the processes involved in the construction of the INEX test collection and describe the resulting components (for a more detailed description, see Fuhr et al, 2003).

3.1. Documents

The document collection consists of the full texts 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[2], it has a suitably complex XML structure (192 different content models in DTD). On average, an article contains 1,532 XML nodes, where the average depth of a node is 6.9.

The overall structure of a typical article consists of a front matter, a body and a back matter. The front matter contains the article's metadata, such as title,

[2] http://trec.nist.gov/

author, publication information, and abstract. The body is structured into sections, sub-sections, and sub-sub-sections. These logical units start with a title, and contain a number of paragraphs tables, figures, item lists, references (citations), etc. The back matter includes a bibliography and further information about the article's authors.

3.2. Topics

The topics were created by the participating groups. We asked each organisation to create a set 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.

The topic format and the topic development procedure were based on TREC guidelines, which were modified to allow for the definition of containment conditions and target elements (e.g. the type of components to return to the user) in CAS queries. The overall structure of an INEX topic consists of the standard title, description, and narrative fields and a new keywords field. Figure 1 shows an example of a CAS topic.

```
<INEX-Topic topic-id="05" query-type="CAS">
  <Title>
    <te>article//tig</te>
    <cw>QBIC</cw>  <ce>bibl</ce>
    <cw>image retrieval</cw>
  </Title>
  <Description>
      Retrieve the title from all articles
      which deal with image retrieval and
      cite the image retrieval system QBIC.
  </Description>
  <Narrative>
      To be relevant a document should deal
      with image retrieval and also should
      contain (at least) one bibliographic
      reference to the retrieval system QBIC.
  </Narrative>
  <Keywords>
      QBIC, IBM, image, video, content query,
      retrieval system
  </Keywords>
</INEX-Topic>
```

Figure 1. A CAS topic from the INEX test collection

The topic development process involved three steps. During the first stage of the topic development process participants created an initial description of their information need without regard to system capabilities or collection peculiarities. During the collection exploration stage participants evaluated, using their own XML retrieval engines, their candidate topics against the document collection. Based on the retrieval results, they then estimated the number of relevant components to the candidate topics. Next, 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.

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 the following criteria: having equal number of CO and CAS topics, having topics that are representative of IR, Database 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.

3.3. Assessments

The final set of topics was distributed back to the participating groups. Participants then used queries generated from any part of the topics, except the narrative, to search the document collection. As a result of the retrieval sessions, participants produced ranked lists of XML elements in answer to a topic. The top 100 result elements from all sixty sets of ranked lists (one per topic) formed the results of one retrieval run. A result element in a retrieval run was identified using a combination of file names and XPath expressions. Associated with a result element were its retrieval rank and/or its relevance status value.

Each group was allowed to submit up to three retrieval runs. We received a total of 51 runs from 25 groups. For each topic, the results from the submissions were merged to form the pool for assessment. The resulting 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 a few journal elements. 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 (listed in alphabetical order) of a given topic, 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. Assessments were provided for 54 of the 60 topics, for a total of 48,849 articles.

4. Evaluation Metrics

Evaluation metrics based on the traditional measures of precision and recall were applied in INEX'02. Based on the framework of (Raghavan et al, 1989), precision was interpreted as the probability, $P(Rel|Retr)$, that a document viewed by a user is relevant. Assuming that the user wants to see NR relevant documents, $P(Rel|Retr)$ is calculated as follows:

$$P(Rel \mid Retr) = \frac{NR}{NR + esl_{NR}} = \frac{NR}{NR + j + \frac{s \cdot i}{r+1}} \qquad (1)$$

The expected search length, esl_{NR}, denotes the total number of non-relevant documents that are estimated to be retrieved until the *NR*-th relevant document is retrieved (Cooper, 1968), and is calculated considering weak ordering (where multiple documents are allowed at a given rank). Given *l* as the rank of the *NR*-th relevant document; *j* is the number of non-relevant documents up to rank *l*-1; *s* is the number of relevant documents to be taken at rank *l* to arrive at *NR* relevant documents; and *r* and *i* are the number of relevant and non-relevant documents at rank *l*, respectively. Raghavan et al. also showed, that intermediary real numbers can be used instead of simple recall points only, hence replacing *NR* in Equation 1 with $x \cdot n$, where *n* is the total number of relevant documents in the collection, and *x* in [0, 1] denotes an arbitrary recall value.

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, based on the set of relevance assessments, *Relevance*:={0, 1, 2, 3}, and the set of coverage assessments, *Coverage*:={N, S, L, E}, quantisation functions were applied on the relevance assessments:

$$f_{quant}(rel, cov) : Relevance \times Coverage \rightarrow [0,1] \qquad (2)$$

With a quantisation function, the overall relevance of a document component could then be determined using the combination of relevance and coverage assessments. A quantisation function can be selected according to the desired user standpoint. For INEX'02, two functions have been selected: f_{strict} and f_{gen}. 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:

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

To credit document components according to their degree of relevance, the quantisation function f_{gen} is used:

$$f_{gen}(rel, cov) := \begin{cases} 1 & \text{if } (rel,cov) = 3\text{E} \\ 0.75 & \text{if } (rel,cov) \in \{2\text{E}, 3\text{L}\} \\ 0.5 & \text{if } (rel,cov) \in \{1\text{E}, 2\text{L}, 2\text{S}\} \\ 0.25 & \text{if } (rel,cov) \in \{1\text{L}, 1\text{S}\} \\ 0 & \text{if } (rel,cov) \in \{0\text{N}\} \end{cases} \qquad (4)$$

Given this (or other) type of quantisation, each document component in a result ranking is assigned a single relevance value. In INEX'02, overlaps of document components in rankings were ignored, thus Raghavan et al's evaluation procedure could be applied directly to the results of the quantisation function. The evaluation results of the applied metrics can be found in (Fuhr et al, 2003).

5. Conclusions and Future Work

In response to the call for participation, issued in March 2002, 49 organisations from 21 countries on four continents registered within six weeks. Throughout the year a number of groups dropped out due to resource requirements, while 6 new groups joined the initiative at the relevance assessments stage. From the 25 groups who submitted retrieval results, 18 attended a workshop, held at the Schloss Dagstuhl Research Centre in Germany[3], which concluded the first round of INEX.

The workshop was organised into presentation and workshop sessions. During the presentation sessions, participants presented their approaches to XML indexing and retrieval. For the workshop sessions three working groups were formed, namely, topic, efficiency and evaluation metrics groups. As a result of the topic working group a new topic format was proposed, based on XPath syntax, which has then provided the basis for the topic format employed in the currently running INEX'03 round. The efficiency working group defined a set of metrics upon which systems can be compared. Based on their recommendations a web site was set up, where data on system capabilities and performance parameters were collected and disseminated to the participants. The workshops on evaluation metrics provided a forum to develop guidelines and procedures for the evaluation of XML retrieval systems based on the employed relevance dimensions. As a result, the proposed evaluation metrics have been implemented and applied to the INEX'02 submissions.

In the second round of INEX[4], running from April 2003 till December 2003, we aim to extend the test collection and develop alternative evaluation measures and metrics addressing the issue of overlapping result elements. One proposed metric, based on the notion of an ideal concept space is currently being developed and applied to the INEX'02 submissions. We are also working on an improved relevance assessment model, which aims to take into account the assessment of fuzzy structural conditions. The two relevance dimensions are currently being re-investigated, in particular the coverage dimension with respect to CAS topics. Finally, we are aiming at ensuring exhaustive and consistent assessments, and we are working on new assessment guidelines, as well as a new online assessment system.

In the long-term future of INEX we aim to extend the range of tasks under investigation to include, interactive retrieval, which will require new evaluation criteria reflecting typical user interaction

[3] www.dagstuhl.de

[4] http://www.is.informatik.uni-duisburg.de/projects/inex03/

with structured documents, and multimedia retrieval, which will make use of XML-based multimedia standards such as MPEG-7.

Acknowledgements

We would like to thank the DELOS Network of Excellence for Digital Libraries[5] for partially funding the first round of the INEX initiative. Special thanks go to the IEEE Computer Society[6] for providing us the XML document collection. Additional acknowledgements go to Deutscher Akadmischer Austausch Dienst (DAAD)[7] and The British Council[8], who supported INEX through their Academic Research Collaboration (ARC) Programme. Last, but by no means least, we would like to thank the participating organisations and people for their contributions to the INEX test collection.

References

R. Baeza-Yates, N. Fuhr, R. Sacks-Davis, and R. Wilkinson, editors. Proceedings of the SIGIR 2000 Workshop on XML and Information Retrieval, 2000.

R. Baeza-Yates, N. Fuhr, and Y.S. Maarek, editors. Proceedings of the SIGIR 2002 Workshop on XML and Information Retrieval, 2002.

H. Blanken, R. Schenkel, T. Grabs and G. Weikum, editors. Intelligent Search on XML, Springer-Verlag (To appear in 2003).

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, 1966.

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.

N. Fuhr, N. Gövert, G. Kazai, and M. Lalmas, editors. INitiative for the Evaluation of XML Retrieval (INEX). Proceedings of the First INEX Workshop. Dagstuhl, Germany, December 8–11, 2002, ERCIM Workshop Proceedings, March 2003.

D. Harman. The TREC conferences. In Hypertext - Information Retrieval - Multimedia, Proceedings HIM ’95, pages 9–28, Konstanz, April 1995.

G. Kazai, N. Gövert, M. Lalmas and N. Fuhr. The INEX Evaluation Initiative. In: Blanken et al. (eds.) Intelligent XML Retrieval. Springer-Verlag (To appear in 2003).

J. Kekäläinen and K. Järvelin. Using graded relevance assessments in IR evaluation. Journal of the American Society for Information Science and Technology, 53(13), 2002.

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.

T. Saracevic. Evaluation of evaluation in information retrieval. In Proceedings of the 18th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, pages 138–146, 1995.

E. M. Voorhees and D. K. Harman, editors. The Tenth Text Retrieval Conference (TREC 2001), Gaithersburg, MD, USA, 2002. NIST.

[5] http://delos-noe.org/
[6] http://computer.org/
[7] http://www.daad.de/
[8] http://www.britishcouncil.org/