Report on the INEX 2003 Workshop, Schloss Dagstuhl, 15-17 December 2003

Norbert Fuhr
University of Duisburg-Essen, Germany.
fuhr@uni-duisburg.de

Mounia Lalmas
Queen Mary University of London, UK.
mounia@dcs.qmul.ac.uk

1 Introduction

The widespread use of the eXtensible Markup Language (XML), especially the increasing use of XML in scientific data repositories, digital libraries and on the web, brought about an explosion in the development of XML retrieval systems to store and access XML content [BGS$^+$03]. These retrieval systems exploit the logical structure of the documents, which is explicitly represented by the XML markup, and retrieve document components (i.e. XML elements) instead of the whole documents. Therefore, XML retrieval systems need not only to find relevant information in the XML documents, but also to determine the appropriate level of granularity to return to the user. In addition, the relevance of a retrieved element is dependent on meeting both content and structural conditions.

Evaluating the effectiveness of XML retrieval systems requires a test collection where the relevance assessments are provided according to a relevance criterion that takes into account the imposed structural aspects. A test collection as such has been built as a result of two rounds of the Initiative for the Evaluation of XML Retrieval (INEX). This initiative provides an opportunity for participants to evaluate their XML retrieval approaches using uniform scoring procedures and a forum for participants to compare their results.

The second round of INEX, INEX 2003, started in March 2003 and ended in December 2003. On December 15-17, 2003, INEX 2003 held its annual workshop in Schloss Dagstuhl, Germany. Around 40 groups registered for participation in INEX 2003, from which 24 groups submitted retrieval runs (several groups had problems with the complex structure of documents and queries and thus were not able to produce retrieval results). The workshop was attended by 45 people from 22 participating groups who presented their results and discussed specific issues of the XML retrieval evaluation methodologies in several working groups. This paper reports on the work presented and discussed at the workshop.

We first describe the INEX testbed. We then give a brief survey over the approaches presented at the INEX workshop. We continue with a summary of the outcomes of the working group. We finish with some conclusions and outlook for INEX 2004.

2 The INEX testbed

The INEX document collection is made up of the full-texts, marked up in XML, of 12,107 articles of the IEEE Computer Society's publications from 12 magazines and 6 transactions, covering the period of 1995–2002, and totalling 494 megabytes in size. The collection 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 (More details can be found in [FGKL03]). Overall, the collection contains over eight millions XML elements of varying granularity (from table entries to paragraphs, sub-sections, sections and articles, each representing a potential answer to a query).

In order to consider the additional functionality introduced by the structure, two types of retrieval queries are used:

- *Content-only* (CO) queries are standard information retrieval (IR) queries similar to those used in TREC. The goal of the IR system is to retrieve the most specific XML element(s) answering the query in a satisfying way. Thus, a system should e.g. not return a complete article where a section or even a paragraph of the same document may also be sufficient. This standpoint is considered in the relevance assessments (see below).

- *Content and structure* (CAS) queries contain conditions referring both to content and structure of the requested answer elements. A query condition may refer to the content of specific elements (e.g. the elements to be returned must contain a section about a particular topic). Furthermore, the query may specify the type of the requested answer elements (e.g. sections should be retrieved). The query language defined for this purpose is a variant of XPath 1.0 [CD99].

Queries were proposed by the participating groups. INEX collected around 120 candidate topics, from which 66 topics were selected to be part of the collection: 36 CO queries and 30 CAS queries.

For the construction of the relevance assessment, INEX employed two relevance dimensions, exhaustivity and specificity, each measured on a multi-grade scale. A given element's degree of relevance combines a measure of how exhaustive it discusses the topic of request and a measure of how focussed it is on the topic of request (i.e. discuss no other irrelevant topics). The assessment procedure made explicit use of the nested XML nested structure to obtain assessments for each level of granularity, that is, both ascendant and descendant elements of a relevant element had to be assessed. As a result, the test collection in INEX consists of nested relevant elements, i.e. sub-trees of the XML articles, where each such element is identified by its absolute XPath expression. Relevance assessments were performed by members of the participating groups, where each group was responsible for about 2 topics. The assessment pools were created by pooling the top 100 results (of 1,500) for each topic from each of the submitted runs. The pools were then assigned for assessment either to the original topic authors or, when this was not possible, on a voluntary basis, to groups with expertise in the topics subject area.

As retrieval measures, recall and precision were used in different variants. These measures were applied both with a binary scale (treating only fully specific and exhaustive answers as being relevant), and in a generalised way where marginally exhaustive/specific answers were counted as fractions of relevant documents. Since recall and precision measures assume implicitly that answers are approximately of the same size and are independent of each other, INEX also uses XML-specific generalisations of these measures considering both the size of answer elements and possible overlap (i.e. when one answer element is contained within another one, the overlap is counted only once).

For the CAS queries, two interpretations were used in evaluation: For the strict view, all structural conditions must be matched strictly (SCAS topics). In contrast, the vague view allows for vague interpretation of the structural conditions (VCAS topics).

3 Results

The participating groups used a broad variety of approaches for performing XML retrieval. Many approaches were based on established IR models like e.g. vector space model, language model, logistic regression or a Bayesian inference model. Others focused more on system aspects, like e.g. adding an XML-specific post-processing step to a “normal” text retrieval engine, using a relational database system for query processing, performing retrieval in a distributed environment. In the following, we give a brief description of the approaches for processing CO queries presented at INEX 2003.

3.1 Model-oriented approaches

Language models
Ogilvie and *Callan* (CMU, USA) extended a standard language model to cater for the hierarchical structure of XML documents. Here the language model of a parent node is computed as the weighted sum of its children's language models, where the weights are proportional to the length of the child node. They tried

various smoothing techniques, but found that context sensitive smoothing was not significantly better than a single collection model. Introducing node type priors for raising the importance of certain elements (e.g. section titles) also had only very little effect

Kamps, de Rijke and *Sigurbjörnsson* (U. Amsterdam, NL) used a multinomial language model with Mercer smoothing. For each element, a separate language model was estimated by linear interpolation of the element-specific model and the collection model. The analysis of relevance judgements from INEX 2002 had shown that relevant elements are larger on average than arbitrary elements. In order to consider this fact, they introduced a length prior. By posing additional constraints on the length and type of elements to be ignored during retrieval, they obtain very good results.

Abolhassani, Fuhr and *Malik* (U. Duisburg-Essen, DE) investigated the application of Amati/Rijsbergen's "retrieval as deviation from randomness" framework. They claimed that this model should be extended for XML retrieval, in order to consider the hierarchical structure of XML documents. For this purpose, they complemented document length normalisation by a factor specifying the hierarchical level of the element under consideration.

Other probabilistic models

Piwowarski, Vu and *Gallinari* (LIP6, Paris, FR) applied Bayesian Networks. As possible element states, they distinguished between the assessments "too big", "exact" and "non-relevant" (used in INEX 2002). For estimating the relevance of a node, the query-node similarity as well as the estimated relevance (state) of the parent node were considered. The corresponding parameters of the link matrix were defined for each element type and then trained on the relevance data from the INEX 2002 collection.

Larson (Berkeley, USA) applied logistic regression and tried component and algorithm fusion by combining the regression results with those from the CORI and BM25 retrieval functions. However, since the regression parameters were not trained on INEX, there is still the opportunity for significant performance gains.

Result Fusion

Ben-Aharon, Cohen, Grumbach, Kanza, Mamou, Sagiv, Sznajder and *Twito.* (Hebrew U., IL) implemented different retrieval strategies considering various aspects of the text and the XML structure, and then performed result fusion for combining the results of these strategies.

Mass and *Mandelbrod* (IBM, Israel) extended the vector space model for XML retrieval. They created separate indexes for each component type to be considered during retrieval, and then performed retrieval runs for each component type. In a subsequent step, results were merged after normalising the scores from the different runs. Overall, this strategy resulted in very good retrieval results.

Enriched representations

Schenkel, Theobald and *Weikum* (MPI Saarbrücken, DE) proposed ontology-based search. For this purpose, they took term-term relationships from WordNet and combined them with statistical weights from large corpora. By adding this information to the original query, retrieval quality could be improved significantly.

Larsen, Lund, Andresen and *Ingwersen* (RSLIS, Copenhagen, DK) investigated the usage of multiple representations. Besides the article text itself, they also considered terms from the INSPEC thesaurus and citation information.

Other models

Doucet, Aunimo, Lehtonen and *Petit* (U. Helsinki, FI) used so-called "maximal frequent sequences" (a sort of statistical phrases) for expanding the original queries. Indexing weights were computed for the leaf nodes of the XML documents, and then propagated to higher level nodes by means of an "augmentation" method where the weights are reduced, in order to identify minimal retrieval units satisfying the query.

Hatano, Kinutani, Watanabe, Mori, Yoshikawa and *Uemura* (Japan) investigated the length of relevant components and came up with a query classification: *SCO* queries contain very specific terms and proper names, and their corresponding relevant components are small; *ACO* queries (aggregate topics) need larger elements as relevant answers. However, they did not yet come up with a system exploiting this finding.

Crouch, Apte and *Bapat* (U. Minnesota / Persistent, USA) adapted the extended vector space model for XML. In this approach, the document vector is split into subvectors, each corresponding to a specific component type of the XML documents.

Liu and *Chu* (UCLA, USA) proposed a cooperative query answering scheme for CAS queries. The system uses various query relaxation techniques and approximate matching of query conditions for answering queries in cooperation with the user.

3.2 System-oriented approaches

Database systems

List, Mihajlovic, de Vries, Ramirez and *Hiemstra* (CWI Amsterdam / U. Twente, NL) presented the TIJAH XML-IR system, a layered database system for XML retrieval. At the logical level, the system uses a probabilistic region algebra (an extension of the "deterministic" region algebra for texts). For processing queries at the physical level, there is a number of "patterns" that lead to specific query execution strategies.

Geva and *Leo-Spork* (QUT, AUS) created an inverted file with path information and stored it in a a relational database system. This way, retrieval could be performed by means of SQL queries. They used coordination level match as retrieval function, considering idf weights only for breaking ties.

Henrich, Robbert and *Lüdecke* (U. Bayreuth, DE) used the Oracle text search component in combination with a stream-based IR approach for combining component weights for different query conditions.

IR systems

Kelly, Geva, Sahama and *Loke* presented a peer-to-peer IR system based on .net software.

Pehcevski, Thom and *Vercoustre* (RMIT, AUS) combined a (Boolean) XML database engine with an IR system for flat documents.

Sauvagnat, Hubert, Boughanem and *Mothe* also used an IR system for flat documents and then post-processed the output in order to account for the XML document structure.

Trotman and *O'Keefe* (U. Otago, NZ) created a new, compressed index structure grouped by element names. Queries were first processed as Boolean queries using bit strings instead of inverted lists, and then the remaining nodes were ranked based on BM25 weights. This strategy led to both good retrieval results and fast query processing.

Flörke (doctronic, DE) developed a proprietary IR engine supporting the notion of "roles" for grouping XML elements with similar semantics. For these roles, prior weights can be defined to be considered during retrieval.

4 Working groups

In addition to the presentations of XML approaches by the workshop attendees, four working groups were formed to discuss issues specific to the evaluation of content-oriented XML retrieval approaches.

Query format

The first working group was on query format, and was chaired by Börkur Sigurbjörnsson from the University of Amsterdam. The main discussion was about the complexity of the INEX 2003 CAS topic format. It seems that people found it difficult to formulate the XPath-like expressions of the topic title (63% of the submitted CAS topics were incorrectly formatted). In view of this high error rate there was discussion about syntax clarification, expressiveness restrictions and even a new syntax for CAS topics. After lengthy discussion, the working group agreed on using only a subset of XPath for structural conditions, extended by vague predicates for specifying element content. The final syntax is yet to be defined. There was also discussion about the difficulty of expressing natural information need. It was questioned whether topic authors added structural constraints because they thought it is useful of whether they did it because they had to write a structured query. To attempt to, at least partly, overcome for unnatural structural information needs, the work group suggested the use of other collections allowing for a larger variety of 'types' of answers.

Relevance
The second working group was on the definition of relevance in XML retrieval, and the relevance assessment procedure. Jaana Kekalanen from the University of Tampere chaired this group. As discussed earlier, in INEX, relevance is defined according to two dimensions, each of them being defined using a multi-graded scale. These definitions were found to be acceptable. It was also acknowledged that these definitions eased the decision of whether an element was relevant or not. The working group had a lengthy debate on what should be the least unit to assess. The relevance of some elements like references (and some other small units) was found to be too difficult to judge. It was agreed that elements that do not need to be retrieved or assessed should be explicitly specified in the future. Points were raised regarding the validity of the assessment of VCAS and SCAS topics, which were done as part of one process. These are currently under investigation. Finally, a number of general suggestions were made: first, to measure retrieval progress by re-using the old topics (INEX'02) with the new version of the systems, and second, to have more topics but without more assessment tasks. This could be achieved by having an easier assessment process (e.g. by defining least judgeable unit) and having more participants. The use of a different type of data (not computer science) was also raised.

Online assessment tool
The third working group was on the online assessment tool, and was chaired by Benjamin Piwowarski, from LIP6, Paris. For carrying out the relevance assessment task, participants used an online tool built by LIP6. The aim of the tool was to ensure exhaustive and consistent assessments of the XML elements with respect to the topics of request. To achieve these, rules were implemented to force assessors to assess all elements that needed to be assessed as they could also be relevant (e.g. if a element was assessed relevant, its parent had also to be assessed), and to check that the assessments were consistent (e.g. the parent element of a relevant element could not be assessed as not relevant). The assessment task is a long and tedious task, and every means to facilitate the task should be investigated while at the same time insuring the quality of assessments, which is crucial for the evaluation results to be meaningful. Existing and additional rules were discussed in details. A common agreement was reached on the rules that will be implemented for INEX'04. A new set of icons that were considered to help in the task was also agreed upon.

Metrics
The final working group, chaired by Gabriella Kazai from Queen Mary University of London, was on metrics. An overview of the current official metrics, inex_eval and inex_eval_ng was given. inex_eval applies the measures of precall as defined in [RBJ89] to document components and computes the probability that a component viewed by a user is relevant. A problem with this metric is that it ignores possible overlaps between result elements and rewards the retrieval of a relevant component regardless if it has already been seen by the user either in full or in part. inex_eval_ng aims to provide a solution to this problem by incorporating component size and overlap within the definition of recall and precision. This metric is based on an interpretation of the relevance dimension within an ideal concept space [WY95]. Additional solutions to solve the overlap of result elements were also proposed; e.g. removing overlapping results from submissions, penalise overlapping result (which was agreed). Alternative metrics were also presented, which attempt to take a more user-oriented view of the relevance of returned elements. Additional quantisation functions (i.e. mapping the two relevance dimensions with their multi-graded scales into single values) were also suggested, to obtain for instance specificity-oriented metrics, and exhaustivity-oriented metrics. Another suggestion was to provide effectiveness results for P@5, P@10, P@20. The working group ended with a discussion on which metrics to be used on which tasks (i.e. CO, SCAS and VCAS).

5 Conclusion and Outlook

The INEX workshop showed that XML retrieval is a challenging new field within IR research. The participating groups are applying a broad range of approaches, most of which are extensions of models developed for "flat" text documents. Since these extensions often include a variety of new tuning parameters, the retrieval results presented so far do not yet allow for judging about the quality/suitability of the different methods.

In addition to learning more about XML retrieval approaches, the workshop allowed participants to contribute to the evaluation methodologies. This has resulted in a number of participating groups to not only continue to develop their XML approaches, but to be actively involved in various aspects of the evaluation tasks (i.e topic format syntax + parser, metrics, etc).

INEX 2004 will start in March of this year, and in addition to the standard ad-hoc task, has 4 new tracks:

- *interactive track* focusing on interactive XML retrieval, considering also navigation through the hierarchical structure,
- *heterogeneous collection track*, comprising various XML collections from different digital libraries, as well as material from other computer science-related resources,
- *relevance feedback track* dealing with relevance feedback methods for XML,
- *natural language track* where natural language formulations of CAS queries have to be answered.

In the future, we are planning to look at XML-based multimedia test-beds, as this is an important issue in many applications such as in those in the context of digital libraries.

Acknowledgements

We would like to thank all the participants for their contributions to and involvement in INEX. In particular, we would like to thank Saadia Malik, Norbert Gövert, Benjamin Piwowarski and Gabriella Kazai for their active involvement in the evaluation tasks and methodologies.

References

[BGS+03] Henk M. Blanken, Torsten Grabs, Hans-Jörg Schek, Ralf Schenkel, and Gerhard Weikum, editors. *Intelligent Search on XML Data, Applications, Languages, Models,Implementations, and Benchmarks*, volume 2818 of *Lecture Notes in Computer Science*. Springer, 2003.

[CD99] J. Clark and S. DeRose. XML path language (XPath) version 1.0. Technical report, November 1999. http://www.w3.org/TR/xpath20/.

[FGKL03] 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, 2003. ERCIM. http://www.ercim.org/publication/ws-proceedings/INEX2002.pdf.

[RBJ89] V.V. Raghavan, P. Bollmann, and G.S. Jung. A critical investigation of recall and precision. *ACM Transactions on Information Systems*, 7(3):205–229, 1989.

[WY95] S.K.M. Wong and Y.Y. Yao. On modeling information retrieval with probabilistic inference. *ACM Transactions on Information Systems*, 13(1):38–68, January 1995.