Accessing XML documents: the INEX initiative

Mounia Lalmas, Thomas Rölleke, Zoltán Szlávik and Tassos Tombros
Department of Computer Science
Queen Mary University of London
Mile End Road, London E1 4NS

Abstract

This document reports on the work carried out during INEX 2004, the evaluation initiative for the evaluation of XML retrieval. We discuss the main track of the initiative, the ad hoc track, and two tracks, the interactive track and the heterogenous track, all of them crucial for investigating more effective, usable and realistic access to XML-based digital libraries.

1 Introduction

As stated in the objective of Workpackage 7[1] digital libraries need to be evaluated to determine how useful, usable, and economical they are and whether they achieve reasonable cost-benefit ratios. Results of evaluation studies can provide strategic guidance for the design and deployment of future systems, can assist in determining whether digital libraries address the appropriate social, cultural, and economic problems, and whether they are as maintainable as possible.

One task of workpackage 7 is the evaluation of content-oriented access to XML documents, where XML stands for "extensible Markup Language", which is increasingly being used in digital libraries and similar systems or plaforms (web, intranet, etc). This increase in storing data/information in XML format brought about an explosion in the development of systems to store and access XML content (three workshops on XML retrieval were held at the SIGIR[2] conference [1, 2, 3]. The aim of such systems is to exploit the logical structure of documents, which is explicitly represented by the XML markup, and access document components, instead of whole documents, in response to a user query. Evaluating how good these systems are, hence, requires

[1] http://dlib.ionio.gr/wp7/aims.html
[2] www.acm.org/sigir/

test beds where the evaluation paradigms are provided according to criteria that take into account the imposed structural aspects.

INEX (Initiative for the evaluation of XML Retrieval[3]) deals with the evaluation of access methods for XML documents. The major tasks involved have been identified as retrieval for content-only queries and retrieval for queries referring both to content and structure of the target elements. During 2004, INEX has undertaken 4 additional tracks dealing with relevance feedback, natural language queries, heterogeneous collections and interactive retrieval in addition to the main track, the so-called ad hoc track. Nearly 60 participating groups have registered for this year's INEX campaign and are currently involved in the relevance assessment task. The results for 2004 will be presented at a workshop in Schloss Dagstuhl, Germany, during 6-8 December 2004 (see [8] for a report of the 2003 INEX workshop).

In this document, we report of the work carried out so far in the ad hoc track, the interactive track and the heterogeneous track. The document finishes with a brief report on the other evaluation work carried out as part of the INEX campaign.

2 Ad hoc track

The aim of the ad hoc track is to evaluate a systems retrieval effectiveness. The ad-hoc retrieval of documents can be defined 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) [9]. 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.

To evaluate ad hoc retrieval, we need a test collection upon which we can measure effectiveness. Creating a test collection requires the selection of an appropriate document collection, the creation of user requests and the generation of relevance assessments.

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 detail can be found in [7]). 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).

[3] http://inex.is.informatik.uni-duisburg.de:2004/

In order to consider the additional functionality introduced by the use of XML query languages, which allows the specification of structural query conditions, INEX defined two types of topics:

- Content-only (CO) queries are standard information retrieval (IR) retrieval similar to those used in TREC. Given such a query, the goal of an XML retrieval 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.
- 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 [6] and was fully specified in [17].

As in TREC [18], an INEX topic consists of the standard title, description and narrative fields. The INEX topics were created by the participating institutions using their own XML retrieval systems or the system provided by the INEX organisers[4] for the collection exploration stage of the topic development process.

Like the topics, the assessments have been derived in a collaborative effort. For each topic, the results from the participants' submissions have been collected into pools using the pooling method [18]. The assessments pools were assigned then to participants; either to the original authors of the topic when this was possible, or on a voluntary basis, to groups with expertise in the topic's subject area. Each group was responsible for about two topics.

The notion of relevance had to be redefined to take into account the nested nature of XML elements. Some elements will be large (e.g. sections) and others small (e.g. paragraphs). Since retrieved elements can be at any level of granularity, an element (the larger element) and one of its child elements (the smaller element) can both be relevant to a given query, but the child element may be more focussed to that given query than its parent element. In this case, the child element is a a better element to retrieve than its parent element, because not only it is relevant to the query, but it is also specific to the query.

The above relates to earlier work on hypermedia document retrieval [4], which showed that the relevance of a structured document can be

[4]In INEX 2003 and 2004, the HyRex system developed in Duisburg-Essen was made available to participants for the topic creation phase, see http://www.is.informatik.uni-duisburg.de/projects/hyrex/index.html.en.

better described by two logical implications. The first one, $d \rightarrow q$ (the document *implies* the query), is the *exhaustivity* of document d for the query q, and models the extent to which the document discusses all the aspects of the query. The second one, $q \rightarrow d$ (the query *implies* the document), is the *specificity* of the document d for the query q, and models to what extent all the aspects of the documents concern the query. Therefore a document d can be exhaustive but not specific to a query, and vice versa. Put in the context of XML retrieval, some XML elements will be exhaustive but not specific to a given query, as they will be too large; whereas some other elements will be specific to a query, but not exhaustive, as they will be too small.

Two dimensions were therefore employed to define relevance:

- Exhaustivity (e-value), measures the extent to which the given element covers or discusses the topic of request.
- Specificity (s-value), measures the extent to which the given element is focused on the topic of request.

Based on this notion of exhaustivity, a section containing two paragraphs, for example, may then be regarded more relevant than either of its paragraphs by themselves. This difference cannot be reflected when using a binary scale for exhaustivity. In INEX, we therefore adopted the following four-point ordinal scale for exhaustivity and specificiy [10]. With respect to exhaustivity:

- Not exhaustive (0), the document component does not discuss the topic of request at all.
- Marginally exhaustive (1)], the document component discusses only few aspects of the topic of request.
- Fairly exhaustive (2), the document component discusses many aspects of the topic of request.
- Highly exhaustive (3), the document component discusses most or all aspects of the topic of request.

With respect to specificity:

- Not specific (0), the topic of request is not a theme of the document component.
- Marginally specific (1), the topic of request is a minor theme of the document component.
- Fairly specific (2), the topic of request is a major theme of the document component.
- Highly specific (3), the topic of request is the only theme of the document component.

To ensure complete assessments, assessors had the use of an on-line assessment system and the task of assessing every relevant document component, and their ascendant and descendant elements within the articles of the result pool [13]. The assessors were given detailed information about the evaluation criteria and about how to perform the assessments. In addition, rules were implemented to ensure consistent assessments (e.g. exhaustivity either increases or remains the same when going from a child element to its parent element).

In INEX 2003, the collected assessments contain a total of 163,306 assessed elements, of which 11,783 are at article level. About 96 % of the 8,802 components that were assessed as highly specific are non-article level elements. This percentage was 87% (of 3,747 components) in INEX 2002. These numbers indicate that sub-components are preferred to whole articles as retrieved units.

To measure effectiveness, we use a metric based on the measure of precall [14] to document components, which computes the probability $P(rel|retr)$ that a component viewed by the user is relevant. That is, it interprets precision as the probability $P(rel|retr)$ that a document viewed by a user is relevant. Given that the user stops viewing at the ranking after he or she has seen NR relevant document components, this probability can be computed as

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

esl_{NR} denotes the expected search length, that is the expected number of non-relevant elements seen in the rank l with the NR-th relevant document plus the number j of non-relevant documents seen in the ranks before (see [5] for details on the derivation). s is the number of relevant documents to be taken from rank l; r, i, are the numbers of relevant and non-relevant elements in rank l, respectively.

[14] give theoretical justification that intermediary real numbers can also be used (here, n is the total number of relevant documents in the collection):

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

This leads to an intuitive method for employing arbitrary fractional numbers x as recall values. The above metric has some theoretical advantages over the more standard recall and precision-based metrics described in [15]: besides the intuitive method for interpolation, it handles ranks containing multiple items correctly.

To apply the metric, the two relevance dimensions are mapped to a single relevance scale by employing a quantisation function. We therefore employ different quantisation to map both dimensions to a single

scalar value. For example, a strict quantisation function is used to evaluate retrieval methods with respect to their capability of retrieving highly exhaustive and highly specific document components. A generalised function is used to credit document components according to their degree of relevance, thus also allowing to reward fairly and marginally relevant elements, i.e. near misses when calculating effectiveness performance.

In INEX 2004, overall performance is obtained by averaging average precision values over all quantisation functions. Full details can be found at http://homepages.cwi.nl/∼arjen/INEX/. In addition, we also show the percentage of overlapping elements per run, as the the metric currently used in INEX does not penalise systems that return overlapping elements (i.e. an element and all its ancestors), which is somewhat defatting the aim of evaluating how XML retrieval systems return the "best" elements.

3 Interactive Track

Issues relating to interactive IR have been extensively investigated in the last decade (e.g. TREC Interactive Track) but these efforts have been mainly in the context of unstructured documents. In the context of XML documents, interactivity is going to different to the one encountered in the conventional case of unstructured documents, and has provided the main motivation for the establishment of an interactive track at INEX [16]. The main aims for the interactive track are twofold. First, to investigate the behaviour of users when interacting with components of XML documents, and second to investigate and develop approaches for XML rerieval which are effective in user-based environments.

The aim of the first year of the interactive track is to investigate the behaviour of searchers when presented with components of XML documents that have a high probability of being relevant. Presently, all metrics that are in use for the evaluation of system effectiveness in the INEX ad hoc track are based on certain assumptions of user behaviour which are not empirically validated. The track also aims to investigate those assumptions.

An experimental methodology was determined and its outline was made available at the track's web site. A baseline interface has been developed, installed, tested and released to participants. A system guide to use the interface has been made available at the track's internal site.

Two kinds of topics have been selected for the interactive track: "Background" type and "Comparison" type topics, and we have selected two topics for each type. This is to see the effects of the topics on user search behaviour.

Questionnaires have been created to question searchers at various stages of the searching procedure: before experiment, before and after each task, after experiment. Instructions for users and experimenter notes have also been created.

The INEX 2004 interactive track guidelines have been written and released to participants. The guidelines describes the motivation for the track, the data collection to be used, the tasks/topics, instructions to be given to searchers, the system to be used, the experimental design, system logs, questionnaires, and the result submission and schedule. Participants may use their own systems to compare with the track's baseline system.

4 Heterogeneous Track

The current INEX collection is based on a single document structure (a single DTD). In practical environments, such a restriction will hold in rare cases only and in particular in digital libraries. It is indeed very likely that in reality most XML collections will comprise documents from different sources, and thus with different DTDs. Also, the increasing use of distributed systems systems (e.g. federations or peer-to-peer systems), will mean that each node (site) manages a different type of collection. The heterogenous track (AKA het track) comes forth from the realization that an information seeker is interested in semantically meaningful answers irrespectively of the structure of the documents, or where it comes from.

The het track aims to answer, among other, the following research questions:

1. For CO queries, what methods are feasible for determining elements that would be reasonable answers? Are pure statistical methods appropriate, or can ontology-based approaches also be helpful?
2. What methods can be used to map structural criteria onto other DTDs?
3. Should mappings focus on element names only, or also deal with element content?
4. What are appropriate evaluation criteria for heterogeneous collections?

In its first year, the track is mainly explorative. The focus is on the construction of an appropriate test collection, and the elaboration of the research issues. For this purpose, the following three tasks were determined:

- creation of a heterogeneous test collection,

- retrieval experiments with a small number of both CO and CAS queries,
- qualitative (rather than quantitative) analysis of the results.

To create a heterogeneous collection, in addition to the INEX collection, six other collections had been selected. These include:

- Berkeley bibliography database,
- Computer Science database of FIZ Karlsruhe,
- University of Duisburg-Essen bibliography database,
- Digital Bibliography & Library Project database,
- Human-Computer Interaction Resources database,
- Publications database of QMUL Department of Computer Science.

We selected ten CO and ten CAS topics from the INEX 2004 ad-hoc topics, and additional topics were created by participants. As one of the tasks of this year, participants documented the process on how they had created the new topics, and we, also, documented the selection process of the ad-hoc topics. Also, guidelines for the topic development in heterogeneous collections was created and made available to the participants. A result submission guideline was released to the participants. It contained the results submission format and procedure, and additional information on XML parsing, where one main issue was how to refer in a unique manner to elements.

5 Others

In addition to the organisation and the running of the ad-hoc task and the four tracks, we are contributing to other development regarding the evaluation of XML retrieval systems.

We are investigating two novel metrics to be used to evaluate XML retrieval systems [12, 19], where user behaviour and intentions are formally encapsulated. The metrics are based on the so-called cumulative gain measures, and the tolerance to irrelevance, respectively, and allow to, for example, represent how much redundant information a user is willing to accept, and how to consider overlapping elements in retrieval runs.

We are also extending the INEX test bed to incorporate multimedia data [11], thus trully reflecting the diverse nature of the data stored in digital libraries. The INEX document collection contains images, and as such can be used to provide a framework where one can evaluate the access to multimedia data within an XML environment.

Acknowledgement

INEX is partly funded by DELOS, a network of excellence in digital libraries. INEX is led by Queen Mary University of London and Duisburg-Essen University. The authors wish to thank the INEX participants for their great involvement in the methodology issues.

References

[1] Baeza-Yates, R., Fuhr, N., Sacks-Davis, R., and Wilkinson, R., Eds. *Proceedings of the SIGIR 2000 Workshop on XML and Information Retrieval* (2000). http://www.haifa.il.ibm.com/sigir00-xml/index.html.

[2] Baeza-Yates, R., Fuhr, N., and Maarek, Y. S., Eds. *Proceedings of the SIGIR 2002 Workshop on XML and Information Retrieval* (2002).

[3] Baeza-Yates, R., and Maarek, Y. S., Eds. *Proceedings of the SIGIR 2004 Workshop on XML and Information Retrieval* (2004).

[4] Chiaramella, Y., Mulhem, P., and Fourel, F. A model for multimedia information retrieval. Tech. rep., FERMI ESPRIT BRA 8134, University of Glasgow, Apr. 1996.

[5] Cooper, W. S. 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* (1968), 30–41.

[6] Clark, J., and DeRose, S. XML path language (XPath) version 1.0. Tech. rep., World Wide Web Consortium, Nov. 1999. http://www.w3.org/TR/xpath20/.

[7] Fuhr, N., Malik, S., and Lalmas, M. Overview of the initiative for the evaluation of xml retrieval (inex) 2 003. In *Proceedings of the Second INEX Workshop* (Schloss Dagstuhl, Germany, March 2004).

[8] Fuhr, N., and Lalmas, M. Report on the inex 2003 workshop, schloss dagstuhl, 15-17 december 2003. *SIGIR FORUM 38*, 1 (June 2004), 42–47.

[9] Harman, D. The TREC conferences. *In Hypertext - Information Retrieval - Multimedia, Proceedings HIM 95.* pages 928, Konstanz, April 1995.

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

[11] Z. Kong and M. Lalmas. Integrating xlink and xpath to retrieve structured multimedia documents in digital libraries. In *RIAO 2004 Conference on Coupling approaches, coupling media and coupling languages for information retrieval*, April 2004.

[12] M. Lalmas G. Kazai and A.P. Vries. The overlap problem in content-oriented XML retrieval evaluation. In *Proceedings of the 27th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, Sheffield, UK, 2004.* ACM, July 2004.

[13] Piwowarski, B., and Lalmas, M. Ensuring consistent and exhaustive relevan ce assessments for xml retrieval evaluation. In *ACM 13th Conference on Information and Knowledge Ma nagement (CIKM)* (Washington DC, USA, November 2004).

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

[15] trec_eval. Evaluation techniques and measures. In Voorhees and Harman [18].

[16] A. Tombros B. Larsen and S. Malik. End users in the context of xml documents: Setting up an interactive track at inex. In *ACM SIGIR Workshop on Information Retrieval in Context*, 2004.

[17] Trotman, A. and Sigurbjrnsson, B. Narrowed Extended XPath I (NEXI). IEX 2004 web site.

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

[19] A.P. Vries, G. Kazai, and M. Lalmas G. Kazai. Tolerance to Irrelevance: A User-effort Oriented Evaluation of Retrieval Systems without Predefined Retrieval Unit. In *RIAO 2004 Conference on Coupling approaches, coupling media and coupling languages for information retrieval*, April 2004.