Report on the INEX 2004 Interactive Track

Anastasios Tombros
Queen Mary University
of London, UK
tassos@dcs.qmul.ac.uk

Saadia Malik
University of Duisburg-Essen,
Germany
malik@is.informatik.uni-duisburg.de

Birger Larsen
Royal School of LIS,
Denmark
blar@db.dk

1 Introduction

As scientific data repositories, digital libraries and publishers increasingly use the eXtensible Markup Language (XML) for publication and storage interest has arisen in exploiting this formatting for retrieval purposes. XML is attractive because it defines the logical structure of the documents and has the potential to assist IR systems in providing more appropriate results to users, i.e., to return relevant document components (i.e. XML elements) rather than whole documents. In addition, the XML tags often have specific semantics that may be exploited purposefully in IR.

Since 2002 the Initiative for the Evaluation of XML Retrieval (INEX) has built test collections where the relevance assessments not only consider whether retrieved components are relevant, but also if they have the right level of granularity. A large range of different XML IR systems and approaches have been tested using traditional laboratory experiments with the INEX test collections [1]. However, relatively little research has been carried out to study user interaction with IR systems that take advantage of the additional features offered by XML documents, and so little is known about how users behave in the context of such IR systems. Therefore, an interactive track was included in INEX 2004 to study these issues. This paper reports on the first year of the track.

We first give the motivation and main aims for the track in section 2, then describe the track's methodology in section 3 and present the main findings in section 4. We conclude and outline the future directions of the track in section 5.

2 Interactive track motivation

In recent years there has been a growing realisation in the IR community that the interaction of searchers with information is an indispensable component of the IR process. As a result, issues relating to interactive IR have been extensively investigated in the last decade. A major advance in research has been made by co-ordinated efforts in the interactive track at TREC (Hersh and Over, 2001). These efforts have been in the context of unstructured documents (e.g. news articles) or in the context of the loosely-defined structure encountered in web pages. XML documents, on the other hand, define a different context, by offering the possibility of navigating within the structure of a single document, as well as of following links to other documents. Studies of user behaviour in such contexts have either been small scale

---

[1] For more information on INEX see *http://inex.is.informatik.uni-duisburg.de/2005/* and the proceedings of INEX 2002-2004 (See Bibliography).

(Finesilver & Reid, 2003), or focused on evaluating a specific interface for XML retrieval (Crestani et al., 2004).

The investigation of the different context that is defined in the case of user interaction with XML documents has provided the main motivation for the establishment of an interactive track at INEX. The aims for the track are twofold. First, to investigate the behaviour of users when interacting with components of XML documents, and secondly to investigate and develop approaches for XML retrieval which are effective in user-based environments.

In the first year, we focused on investigating the behaviour of searchers when presented with components of XML documents that have a high probability of being relevant (as estimated by an XML IR system). Presently, metrics that are used for the evaluation of system effectiveness in the INEX ad-hoc track are based on certain assumptions of user behaviour (Kazai, 2004). These metrics attempt to quantify the effectiveness of IR systems at pointing searchers to relevant components of documents. Some of the assumptions behind the metrics include that users would browse through retrieved components in a linear order, that they would "jump" with a given probability from one XML element to another within the same document's structure, that they would not make use of links to another document, etc. These assumptions have not been formally investigated in the context of XML retrieval; their investigation formed the primary aim for the first year of the interactive track.

More specifically, we wanted to study the way that searchers access other XML elements of a given XML document once they have selected an initial element from the ranked list. One of the issues we were interested in looking into was the order in which searchers browse through the tree that corresponds to the logical structure of a document, and how this order correlates to the relevance assessments that searchers provide for XML elements. From this information we can also make observations about how and when searchers browse to more specific or more exhaustive components within a document.

Since the investigation of user behaviour forms our primary focus, the format of the track for the first year differs to that typically followed by, for example, the interactive track at TREC. The main difference was that a comparison between different interactive approaches was not our main focus. Instead, a more collaborative effort was planned, with the outcome of the studies expected to feed back to the INEX initiative. Participating sites still had the option to develop and evaluate their own interactive approaches, but this was not a requirement for participation. It should be noted that none of the participating sites opted to develop their own system.

3   Methodology

3.1   Topics

We used content only (CO) topics from the INEX 2004 collection. In order to make the tasks comprehensible by other than the topic author, it was required that all INEX 2004 topics not only detail *what* is being sought for, but also *why* this is wanted, and in what context the information need has arisen. Thereby the INEX topics are in effect simulated work task situations as developed by Borlund (2000).

We also selected topics that corresponded to different types of search tasks. In Tombros et al. (2005) it was shown that different task types promote the use of different criteria when assessing the relevance of

web pages. We wanted to investigate whether a similar effect, in terms of user behaviour within structured documents, exists in the context of XML documents.

Four of the 2004 CO topics were used in the study, and they were divided into two task categories: *Background category (B)*, where the topics express an information need in the form of "I'd like to find out about X", and *Comparison category (C)*, where the topics' subject is along the lines of: "Find differences between X and Y".

3.2   System

A system for the interactive track study was provided by the track organisers. The system was based on the HyREX[2] retrieval engine, and included a web-based interface with a basic functionality.

In response to a user query, HyRex returns a ranked list of components, including for each component the title and authors of the document in which the component occurs, the component's retrieval value and the XPath of the component. Searchers can explore the ranked list of components, and can visit components by clicking on the Xpath in the ranked list. In Figure 1 we show the detailed component view, where the actual textual contents of the selected component (right hand) and the table of contents for the document containing the component (left hand) are shown. Searchers can access other components within the same document either by using the table of contents, or by using next and previous buttons provided by the interface. The granularity of all components on the interface is at the article, section, subsection and sub-subsection levels.

A relevance assessment for each viewed component could be given, as shown in Figure 1. The assessment was based on two dimensions of relevance: how useful and how specific the component was in relation to the search task. These dimensions were described to searchers in a way that was both intuitive and that corresponded to the relevance dimensions of the ad-hoc track of INEX. Each dimension had three grades of relevance, (*Very*, *Fairly* and *Marginally*) and ten possible combinations of these dimensions could be made as shown in Figure 1.
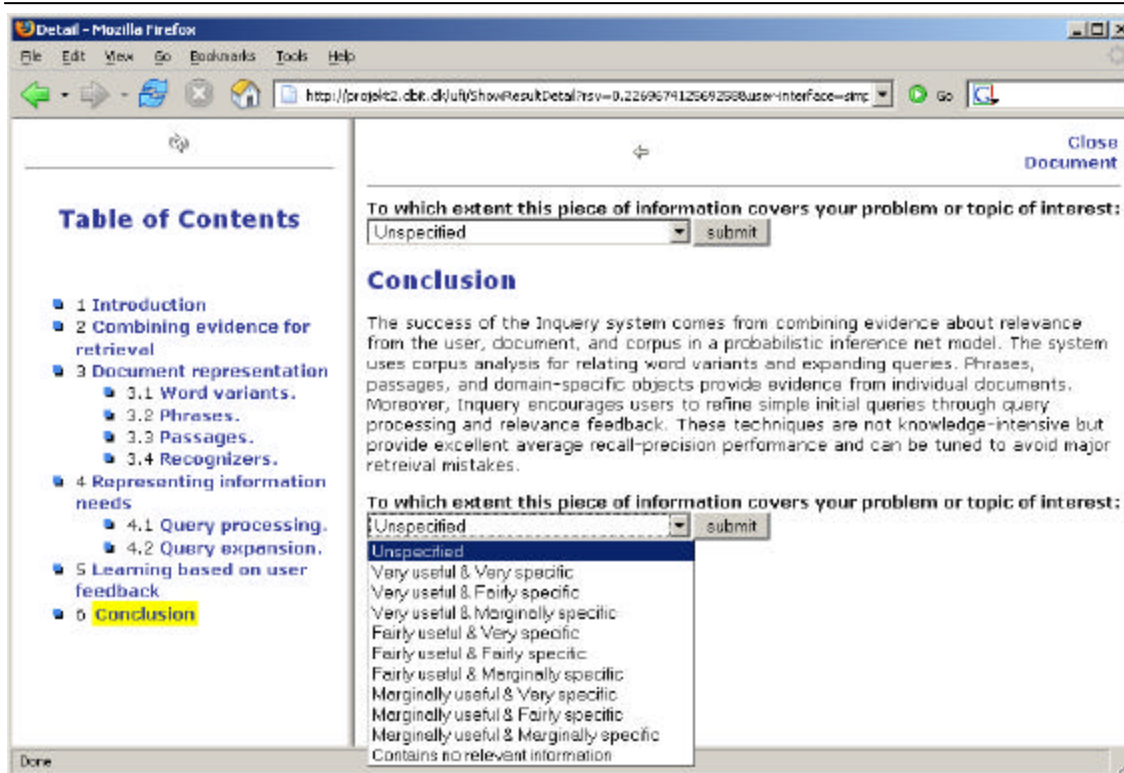
---

[2] http://www.is.informatik.uni-duisburg.de/projects/hyrex/

Figure 1. Detailed component view containing the full text of the component and a table of contents for the whole document, and showing the relevance assessment scale.

3.3 Participating sites

The minimum requirement for sites to participate in this year's interactive track was to provide runs using 8 searchers on the Baseline version of the XML retrieval system that was provided. In addition to the minimum requirement, sites could choose to employ more users, to compare the baseline system with a graphical version, or to test their own experimental system against the baseline system provided. Ten sites participated in the Interactive track (Table 1).

Kyunpook National University, Korea	Robert Gordon University, Scotland
Norwegian University of Science and Technology	Royal School of LIS, Denmark
Oslo University College, Norway	University of Duisburg-Essen, Germany
Queen Mary University London	University of Tampere, Finland
RMIT, Australia	University of Twente/CWI, The Netherlands

Table 1. Participating sites in the Interactive Track at INEX 2004.

3.4 Experimental protocol & data collection

A minimum of 8 searchers from each participating site were used. Each searcher searched on one task from each task category. The task was chosen by the searcher, and the order of task categories was permuted.

Searchers had to fill in questionnaires at various points in the study (before the start of the experiment, before each task, after each task, and at the end of the experiment), and an informal interview and debriefing of the subjects occurred at the end of each experiment. The goal for each searcher was to locate sufficient information towards completing a task, in a maximum of 30 minutes per task.

The collected data comprised questionnaires completed by the test persons, the logs of searcher interaction with the system, the notes experimenters kept during the sessions and the informal feedback provided by searchers at the end of the sessions. The logged data consisted of the queries issued, the components actually viewed and the order in which they were viewed, relevance assessments of these, any browsing behaviour, as well as time stamps for each interaction between searchers and the system.

4   Main findings

A first analysis of the main results of the track was presented in (Tombros et al., 2005b); here we give an overview of the major findings.

Limited amount of interaction with retrieved components. A general observation was that searchers did not interact much with other components of a given XML document after the selection of the initial component from the ranked list. In a large number of cases, they merely visited this initial component and returned to the ranked list. Further, in a large number of cases in this one-component interactions, the assessments provided for the accessed component were rather low. This limited interaction prompted us to further look into the reasons that might have caused it.

Component overlap. One of the possible reasons was the strong presence of component overlap in the retrieved list of elements. Overlapping components are components from the same document, retrieved at different ranks in the hit list. The HyREX retrieval engine as used in the track did not deal with overlapping components, and therefore searchers frequently ended up accessing components of the same document from the hit list.

Data from both the system logs and the questionnaires suggested that searchers found the presence of overlapping components distracting. By recognising that they have accessed the same document through a different retrieved component, searchers would normally return to the hit list and access a different retrieved component instead of browsing again within an already visited document. This behaviour contributed to a large number of the one-component interactions mentioned previously.

Component summaries. A further contributing reason to the one-component interaction was the presentation of the retrieved elements in the hit list. The hit list presentation in the system used in this study did not include any kind of XML document, or element, summarisation; only the title and authors of the document were displayed in addition to the XPath of the component and its similarity to the query. This implied that searchers had little clues available to decide on the usefulness or not of retrieved elements. The lack of available clues is one of the main reasons that led searchers to give low assessment scores to the first component they visited from the ranked list.

Document structure provides context. The presence of the logical structure of the documents alongside the contents of the accessed components was a feature that searchers commented positively on. The Table of Contents of each document (Figure 1) seemed to provide sufficient context to searchers in order to decide on the usefulness of the document or not. The specific nature of the documents of the INEX collection (scientific articles) could have also assist searchers in extracting extra context form the logical

structure of documents: the idiosyncratic nature of scientific articles allows searchers to expect specific rhetorical roles to be fulfilled in specific parts of a document (e.g. Introduction, Methodology, Conclusions, etc.). It would certainly be worthwhile to investigate whether similar observations would hold when different document types are used.

Browsing within XML documents. An initial analysis of the detailed browsing behaviour of searchers looked into whether searchers tended to browse to more specific information (e.g. from a section to a subsection) or to more exhaustive (or general) information (e.g. from a subsection to a section). The results indicate that much more often searchers browse to components of the same granularity (e.g. from a subsection to another subsection) than to components of different granularity. Searchers also tend to browse to more specific information more often than they do to more exhaustive information. These results also seem to correlate with relevance assessments given by searchers. For example, searchers are more likely to browse to a component of the same granularity after they have assessed a component as very useful or very specific, than after having given other assessment types. We are currently analysing data in this category in more detail.

Other issues. In relation to differences in searcher behaviour when performing tasks of different types (background or comparison), few differences were noted. A contributing reason to this was the lack of diversity among the INEX 2004 CO topics. Most of the topics fall under the general *Background* category, and even the ones we selected for the *Comparison* category are not as distinct from that category as we would have hoped. We plan to further look into differences for different task types with more diverse topics in future runs of the track.

Regarding whether searchers preferred XML components, as opposed to full articles, we did not have enough evidence to conclude. Searchers seemed to have a similar preference to accessing full articles and components at the section level from the hit list. They also seemed to have a higher preference for components at the section level than at the finer granularity subsection levels when they browsed within XML documents. We plan to look into this issue more carefully in INEX 2005.

5   Conclusions and future directions

In the first year of the interactive track at INEX we have gained some basic knowledge about how users interact when presented with document components rather than whole documents. The future focus of the track will be on methods to support users by means of features derived from the XML formatting.

Overlapping components, i.e., components from the same document at different ranks in the hit list, frustrated many users but might instead be exploited to achieve a better and more comprehensible presentation of results, e.g., by hierarchical hit lists or for highlighting parts of documents. An issue to be investigated is how to present components in the hit list, so that users get a good impression of whether the component might be worth examining further or not. The support needed for different task types might be different and an attempt to define more different task categories will be made. In addition, it is an open question how to make flexible and advanced querying available that can exploit the XML formatting and at the same time remain sufficiently simple for users to comprehend and use. For INEX 2005, we plan to make a system available for all participants in order to study some of these issues. At the same time, we encourage research groups to develop their own systems to study a broader range of questions.

6   Acknowledgments

7   References

Borlund, P. (2000). Evaluation of interactive information retrieval systems. Åbo: Åbo Akademi University Press. vi, 276 p. (PhD dissertation).

Crestani, F., Vegas, J. and de la Fuente, P. (2004). A graphical user interface for the retrieval of hierarchically structured documents. *Information Processing & Management*, 40(2):269-289.

Finesilver, K. and Reid, J. (2003). User Behaviour in the Context of Structured Documents. In Advances in Information Retrieval: 25th European Conference on IR Research, ECIR 2003, pp 104-119.

Hersh, W., and Over, P. (2001): Interactivity at the Text Retrieval Conference (TREC). *Information Processing & Management*, 37(3), 365-367. (Introduction to special issue)

Kazai, G. (2003). Report of the INEX 2003 metrics working group. In Proceedings of the 2nd Workshop of the Initiative for the Evaluation of XML retrieval (INEX), pp 184-190.

Tombros, A., Ruthven, I. and Jose, J. (2005). How users assess web pages for information seeking. Journal of the American Society for Information Science and Technology, 56(4):327-344.

Tombros, A., Larsen, B. and Malik, S. (2005b). The interactive track at INEX 2004. In Proceedings of the 3rd Workshop of the Initiative for the Evaluation of XML retrieval (INEX).

INEX Proceedings
INEX 2002 Workshop Proceedings (Available: *http://www.ercim.org/publication/ws-proceedings/INEX2002.pdf*)

INEX 2003 Workshop Proceedings (Available: *http://inex.is.informatik.uni-duisburg.de:2003/proceedings.pdf*)

INEX 2004 Workshop Proceedings (Forthcoming as LNCS 3493, Pre-proceedings available: http://inex.is.informatik.uni-duisburg.de:2004/pdf/INEX2004PreProceedings.pdf)