User Expectations from XML Element Retrieval

Stamatina Betsi[1], Mounia Lalmas[2], Anastasios Tombros[2] and Theodora Tsikrika[2]

[1] LogicDIS S.A., Greece
s.betsi@gmail.com

[2] Queen Mary, University of London, UK
{mounia,tassos,theodora}@dcs.qmul.ac.uk

ABSTRACT

The primary aim of XML element retrieval is to return to users XML elements, rather than whole documents. This poster describes a small study, in which we elicited users' expectations, i.e. their anticipated experience, when interacting with an XML retrieval system, as compared to a traditional 'flat' document retrieval system.

Categories and Subject Descriptors

H.3.3 [**Information Search and Retrieval**]

General Terms

Human Factors, Experimentation

Keywords

User study, XML, element retrieval

1. INTRODUCTION

The primary aim of XML (eXtensible Mark-up Language) element retrieval is to return to users XML elements, rather than whole documents, in order to reduce their effort when viewing large texts, by allowing them to focus only on the parts of the document that are relevant to their information needs. XML retrieval is a relatively new research field and recent research has played a part in the effort to establish the goals of XML retrieval systems [3]. Studies on interaction aspects, however, are still limited (e.g. [5]). We believe that it is important to ask real users to describe what they believe would enhance their interaction experience with an XML retrieval system, and to express their concerns, needs, and desires. Our aim is to elicit users' expectations regarding interaction in the context of XML retrieval in order to show whether and possibly how XML retrieval can fulfil these expectations. Here, the term "expectations" is used to describe what a user anticipates to experience by using such a system.

2. METHODOLOGY

We elicited user expectations through a series of semi-structured interviews, where the emphasis was on the aspects that differentiate XML retrieval systems from traditional information retrieval (IR).

We interviewed 10 IT employees working in a software development company - LogicDIS (http://www.logicdis.gr). All of them regard retrieving information essential for successfully completing their daily tasks, and their jobs involve extensive use, on a regular basis, of an existing IR system, the LogicDIS Portal, a standard, web-based, filing and retrieval system. The document collection is an internal, private collection consisting of manuals and requirements, design, analysis, test, etc. documents.

SIGIR'06, August 6–11, 2006, Seattle, Washington, USA.
ACM 1-59593-369-7/06/0008.

To perform our study, we needed to highlight the main differences between traditional IR and XML IR. The interviews were organised around interaction units (starting point, query specification, interacting with results, information seeking as a whole) defined following Bates' berry picking information seeking model [1]. Full details can be found in [2].

Throughout the interviewing process, we avoided the use of IR-related terms (including the terms 'XML' or 'XML retrieval'), since due to the users' substantial experience and knowledge in computer science this information could be misinterpreted. Interviews took place in front of the available traditional IR system (LogicDIS Portal) and we encouraged users to describe its advantages and disadvantages. We then presented them, orally or by using paper visual aids, with solutions that an XML retrieval system could possibly provide or the additional issues it might introduce. The main question investigated in this poster was what users/interviewees thought of these solutions.

3. FINDINGS

All users liked the idea of being able to access directly the document parts that they were interested in. This clearly demonstrates that element retrieval, as opposed to document retrieval, has merits. Interviewees, however, expressed a strong wish in "maintaining some control".

A main issue was the presentation of retrieved elements. Almost all users stated or implied that they would expect the retrieved elements to be grouped per document, as they wanted to still "see what the document is". Most users insisted that they preferred this grouping method to a flat, serial presentation of relevant elements. This finding strongly motivates the grouping of elements per document (Fetch & Browse retrieval task) that is investigated in INEX, the evaluation initiative for XML retrieval [3].

In addition to the above grouping, providing information about the relationship between retrieved elements and their parent / children elements was expected so as to avoid becoming disorientated. This was because, as stated by the interviewees, by choosing to view a specific element, in most cases, they do not (necessarily) wish to explicitly view its content, but to interact with this specific piece of information in its context, as the context of information determines several aspects of the information itself.

It was also mentioned that providing some additional information per element would help users to decide which elements are relevant to their information need. In general, a concise description of the content of each retrieved element and its structural type were expected, even in the form of a short title or summary.

A consequence of the above expectations was that users did not consider the presentation of overlapping elements to be an issue [4]. Overlapping results in XML retrieval lead to the presentation of nested document components (e.g. both a

section and one of its paragraphs) as results to the user, who may access them both either directly or through browsing. In this way, redundant information is provided to the user. The expected grouping of the relevant elements within a document, *and* the expected explicit representation of relationships between elements (e.g. indented representation), as suggested or implied by most users, appear to deal with this issue.

A second main issue regarded the types of queries. The existence of a logical structure in XML documents makes it possible to query with respect to content and structure criteria. In XML retrieval there are two types of queries: content-only (CO), where structural constrains are not taken into account and the aim is to return relevant elements at the right level of granularity; and content-and-structure (CAS), where structural constraints are explicitly stated in the query and they can refer both to where to look for the relevant elements and what type of elements to return. Some interviewees suggested the inclusion of a very simple grammar that could be used to identify whether a user wishes for whole relevant documents to be retrieved, or any relevant elements, or relevant elements of a specific type. However, most users stated that it would be difficult, or even impossible, to define in a query the structural part that contains the required information. Furthermore, most users stated that they would prefer all relevant information to be included in the results, regardless of the type of the structural parts that contain it. This seems to correspond to the view adopted in INEX for dealing with the structural constraints of CAS queries, i.e. as hints for where to look and what to retrieve (e.g. [3]).

4. ANALYSIS

This section reports our analysis of the findings, as well as some suggestions regarding the design of XML retrieval systems.

It was clear that users expect to interact with *documents*. Interview results showed that users expect the retrieved components to be accompanied by the documents that contain them. They would feel rather uncertain if elements with no context information were retrieved. There may be several reasons for this. An obvious one is that users (our interviewees) are used to dealing with whole documents. It is also the case, at least in our study, that the documents themselves contained essential "meta-information" about the retrieved information, which should be made visible to users, who felt this could only be done through having access to the whole document.

Users stated that they disliked long lists of retrieved elements. However, an effective XML retrieval system is expected to return a significantly greater number of results than an equally effective traditional IR system, but many of these will be overlapping elements. Presenting the elements as *indented* details of the document, with each element indented according to its level, appears to address the overlapping issue [4], as well as providing users with the expected grouping of elements per document and explicit relationships between elements. Users are informed that the retrieved elements that contain relevant information are nested; hence, they become instantly aware that a parent element contains both its own information and that of its children elements. Hierarchies are a natural structure, frequently used in computers, and extremely familiar to users.

Our analysis of the interviews also suggests that, upon retrieval, users should be able to select any of the presented elements, including the document itself. If the user selects directly an element, then the focus should be on that specific element. If the user selects the document, then the focus could be on the (first) most relevant element in that document. Furthermore, also from the interviews, users did not necessarily expect to be provided with relevant information when they interact with the retrieval results. More importantly, they expect to be provided with information that indicates whether a result contains relevant information or not.

It was mentioned several times by interviewees that relevant information is scattered across many documents. In traditional IR, retrieved items are documents. The users can retrieve a relevant document and save it for later use. This is not so straightforward in the context of XML retrieval. If users save an XML document as it is, then they will lose information about which elements were relevant. If users are provided with capabilities to save only one retrieved element, then they will lose its relationship to its parent element and the document. A solution is to provide users with the capability to collect several elements and store them together with a link to the document, information about the element's relevance, and its relationship with the other elements of that document.

5. CONCLUSIONS

This study aimed at eliciting from real and potential users their expectations from an XML retrieval system compared to a traditional IR system. A number of findings have been reported, many of which confirm what many researchers (in INEX) already speculated to be the case (for instance, how to treat CAS queries). One limitation of this study is the sample of the users participating in this study may not be representative. Furthermore, interviewees were often requested to visualise the functionality of an XML retrieval system, as they were only presented with paper visual aids, and therefore needed to imagine whether or not some functionality would be helpful, or whether it would cause confusion. Further studies employing users with more diverse backgrounds and possibly functional XML retrieval systems are required.

In conclusion, we believe that the presented findings provide a good starting point for further research, as well as a valid reference for the issues that should be taken into consideration when designing XML retrieval systems.

6. REFERENCES

[1] M.J. Bates. The Design of Browsing and Berry-picking Techniques for the Online Search Interface, *Online Review*, 13(5):407-424, 1989.

[2] S. Betsi. *XML Retrieval: Users' Expectations*, MSc in Information Management. Queen Mary, University of London, August 2005.

[3] N. Fuhr, M. Lalmas, S. Malik and G. Kazai (eds). *Advances in XML Information Retrieval* and evaluation *INEX 2005 Workshop Proceedings*. Lecture Notes in Computer Science, vol. 3977. Springer Verlag, 2006.

[4] G. Kazai, M. Lalmas and A.P. de Vries. The Overlap Problem in Content-Oriented XML Retrieval Evaluation. In *Proceedings of the 27th ACM SIGIR Conference*, Sheffield, UK, pp 72-59, 2004.

[5] A. Tombros, B. Larsen and S. Malik. The Interactive Track at INEX 2004. *Advances in XML Information Retrieval*, *INEX 2004 Workshop Proceedings*, LNCS 3943, pp 410-423, 2005.