Obtrusiveness and relevance assessment in interactive XML IR experiments

Birger Larsen
Department of Information Studies
Royal School of Library and Information Science
Copenhagen, Denmark
blar@db.dk

Anastasios Tombros
Department of Computer Science
Queen Mary University of London
London, United Kingdom
tassos@dcs.qmul.ac.uk

Saadia Malik
Fak 5/IIS, Information Systems
University of Duisburg-Essen
Duisburg, Germany
malik@is.informatik.uni-duisburg.de

ABSTRACT
Ensuring realism in Information Retrieval (IR) experiments (whether laboratory or user based) is always a difficult problem. Obtaining relevance assessments of high quality is of pivotal importance to most studies and a significant challenge. In element retrieval from structured documents, where both whole documents but also parts of documents (elements) may be retrieved as answers, the type of research questions being posed accentuates this problem. In this opinion paper we reflect on the range of aspects we would ideally like to have assessed – in particular with regard to involvement of end-users. The problems involved in requiring assessment of several aspects for each interaction are discussed and a number of alternatives considered.

1. BACKGROUND
Documents formatted in XML and similar mark-up languages are attractive for IR because the mark-up 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 can have specific semantics that may be exploited purposefully in IR.

This has formed the impetus behind the establishment of INEX – the INitiative for the Evaluation of XML Retrieval. Since 2002 INEX has built test collections to make it possible to test different XML IR approaches [3]. The central research issue is how to exploit the logical structure of documents (explicitly represented by the XML mark-up) to provide more precise answers to end-users. Therefore the relevance assessments not only consider whether retrieved elements are relevant, but also if they have an appropriate level of granularity. Two important, and logical, extensions to traditional IR have been made to facilitate this: In response to earlier criticism against the limited realism of binary relevance assessments [see, e.g., 9] *graded assessments* are used in INEX to express the degree to which a given element is relevant to the information need, and two *different dimensions of relevance* are considered: exhaustiveness and specificity[1]. While the measurement of performance with these assessments has been facilitated by novel non-binary measures [5-7], the use of graded relevance and the two dimensions continue to be debated in INEX: *First*, as the assessments are provided by humans there are concerns about the consistency of them, in particular with such an elaborate two-dimensional relevance scale. *Second*, as not only the retrieved elements but also their descendants and ascendants need to be assessed, the assessment process becomes very laborious when two dimensions of relevance have to be assessed on graded scales.

[1] *Exhaustiveness* describes the extent to which the component discusses the topic of request, and *specificity* the extent to which the component focuses on the topic of request.

2. REALISM AND ASSESSMENT
From 2004 INEX includes an interactive track. Where the main ad hoc track in INEX facilitates laboratory tests of the performance of different XML IR techniques, the interactive track aims at investigating the behaviour of users when interacting with elements of XML documents, and ultimately to facilitate the development of approaches for XML retrieval which are effective in user-based environments. The interactive track thus attempts to put the techniques developed in the ad hoc track into practise so that they may be used by end-users in realistic search environments. An additional purpose of the track is to give useful information to the main track in INEX. Details about the track and results of an initial analysis of the collected data can be found in [10].

For the first year, the interaction of end-users with an XML IR prototype system was studied. The main goal was to investigate if end users would at all like to have elements as answers (rather than the usual whole documents), how they would browse within documents, and which kinds of elements they would assess as relevant. In order to study this in detail, some sort of relevance assessments were needed. Ideally, we would like to have a number of aspects assessed each time a test person has looked at an element:

1. The amount of relevant information the element contains versus irrelevant information (~ *specificity*),
2. How much of the information need can be solved by the element (~ *exhaustiveness*),
3. Whether the retrieved information is *redundant* or not (i.e., has been seen already in other elements)
4. How *useful* the element is overall in solving the information need.

Together with the sequence of interactions, such detailed information on each viewed element could help answer a number of pressing research questions in XML IR, e.g.,

1. What granularity of retrieved elements do users prefer?
2. What do users gain by browsing up/down the XML tree?
3. Would users rather skim larger parts of documents than risk having smaller irrelevant elements?
4. Are users very sensitive to redundant information?
5. …and ultimately, is the retrieval of elements of value to end users, or would they rather just have the full documents?

However, the cognitive load on the test persons would be great if they had to judge and balance all four aspects for each interaction. Experimentally this is undesirable as it is a goal to minimise the cognitive load deriving from factors that do not occur in normal searching behaviour. Having to interrupt the search to give complex relevance assessments may not only result in an unrealistic searching behaviour, but may even be experienced as *obtrusive* by the test persons. This problem is particularly pressing in XML IR where users are likely to browse the document structure to identify other relevant elements than those initially proposed by the system. We would preferably have the test persons to assess the four aspects for each viewed element, and to ensure capturing this information perhaps even forcing the user to do so before moving on to the next element. This has been tried successfully in IR previously in the Okapi experiments, but with much simpler document surrogates and binary relevance assessments [1]. Asking or forcing user to perform complex assessments on all four aspects would inhibit the natural interaction with the system given the much more complex documents and desired aspect to be assessed in XML IR. Here the risk is that the better part of the test persons' attention would be spent on doing the assessments, and not on the interaction.

A compromise between the ideal situation outlined above and a slightly less obtrusive setting was attempted in the interactive track in 2004. The graded scales and two relevance dimensions from the ad hoc track were maintained[2], but merged into to a single dropdown list with 10 points. Figure 1 displays a screenshot of the system interface including the relevance scale. The prototype system retrieved a ranked list of XML elements. Any element chosen for display was placed in the context of the containing document by showing its position in a table of contents. To allow the test persons to interact as naturally as possible, they were free to choose any element from the ranked list and to browse within the documents as they saw fit. The test persons were, however, instructed to assess viewed elements, but not forced to.

However, this method of collecting assessments also presented some drawbacks. It did not guarantee, for example, that test persons would provide assessments for every single element they viewed; it was possible for them to leave a viewed page without providing any assessments. Unassessed elements were viewed as providing an indication of non-relevance. However, there is no tangible evidence to suggest that this is always the case. Further, although the test persons provided a quantitative indication of relevance, they did not provide a qualitative one, i.e., why was a certain element too specific or too exhaustive, or why was a certain element not relevant at all? This kind of qualitative data was not captured explicitly in the experimental set up, but was mostly inferred by the logs of the search sessions, the time stamp data, etc.

Very few of the test persons communicated difficulties in understanding or using the 10 point relevance scale. Nevertheless, the results of an initial analysis of the collected assessments indicate that the test persons may have had such difficulties as parts of the scale were underused [8]. In addition, only 60% of the viewed components were assessed [10] and there are qualitative comments in the questionnaire data indicating that some test persons were tired of having to assess every viewed element.

The next section lists a number of alternatives and discusses the advantages and limitations of each.

3. ALTERNATIVES

A first alternative would be *not* to ask the test persons to assess the documents at all, and use the relevance assessments from the ad hoc track instead. This approach would provide easy access to already available relevance assessments, and would impose minimum strain on the test persons. On the other hand, such an approach is fundamentally opposed to the very idea of interaction and of simulated work-task situations; the subjective notion of relevance is disregarded in this approach.

A second alternative would be to use implicit indicators of relevance (as opposed to explicitly indicating relevance by means of a quantitative scale). Implicit indicators can include the time spent viewing an element, the amount of scrolling involved, etc. [12]. This approach would also impose a minimum strain on the test persons as indicators of relevance would stem from the way the persons interact with elements. However, an inherent difficulty with relating relevance to implicit indicators, is that there is no unambiguous evidence that the behaviour indeed suggests relevance. For example, a test person may choose to spend longer time reading the contents of an element because he finds this element difficult to comprehend, and not necessarily because he finds it relevant. Further, it is also difficult to correlate implicit indicators with certain levels of exhaustiveness or specificity.

Considering some more concrete indicators of relevance based on user behaviour is also possible. For example, test persons may be asked to bookmark elements that are of interest. This approach is also cognitively easy on test persons, as the act of bookmarking is rather natural during information seeking tasks nowadays. In addition, it allows us to consider in detail fewer elements for further analysis, i.e. to focus on the elements that test persons found interesting. However, this approach would present us with a large fraction of viewed elements for which no data is available.

Alternatively, if every viewed element is to be assessed the act of doing so should be made as straightforward and easily comprehensible to the test persons as possible because of the associated cognitive load and risk of obtrusiveness. A complex relevance scale, such as those used in the ad hoc and 2004 interactive track, or the assessment of several aspects for each interaction work against this. Rather a simple scale gauging a

[2] *Exhaustiveness* was renamed *Usefulness*, but the same definition was used in the instructions for test persons.

single aspect or one dimension should be employed. A simple scale will allow the test persons to complete their assessments without much delay, and have been successfully implemented in interactive IR experiments in the past (See e.g. [2]). A limitation both with bookmarking and the use of a simple scale is that data about why and how each element is relevant would still not be made available by the test persons.

It is possible to obtain explicit accounts of why elements were assessed at a certain relevance level through the use of more sophisticated equipment and experimental techniques. For example, it is possible to use eye tracking equipment to monitor the test persons' eye movements while reading the contents of elements. By analysing fixation periods and saccades, it is possible to make inferences about the test persons' perception of importance of the various elements. This can be combined with a structured interview after the search session, in which the test persons will elaborate during a replay of the session why certain decisions were made [4]. Such 'talk-after interviews' can also be carried out with less expensive on-screen video capturing software [11]. Alternatively, think-alouds could also be employed during the search sessions in order to capture the reasons for the test persons' assessments. These approaches have the advantage that they enable us to document why test persons assess certain elements at a specific relevance level. However, the need for specialised equipment and for more laborious experimental techniques (e.g. analysing structured interviews) may present some practical challenges in implementing this approach.

4. CONCLUSIONS

In order to answer some of the important fundamental questions in XML IR, a wide range of aspects should ideally be assessed at each interaction with the test system. This would, however, prevent the tests persons from interacting naturally with the system, and thus undermine the purpose of an interactive study.

Therefore, different alternatives were discussed. A common thread in these is the challenge of finding a method that can inform us why something is relevant or not-relevant, while at the same time not being obtrusive enough to obscure the browsing and searching behaviour of the test persons.

We do not regard complex relevance scales such as those employed in the ad hoc track or a requirement to judge several aspects for each viewed element as fruitful in an interactive setting. Instead, bookmarking or the use of a simple scale should be used to minimise the cognitive load on the test persons and allow a searching behaviour that is as natural as possible. This in combination with eye-tracking or desktop video approaches may help answer some of the important research questions in XML IR by allowing the collection of data that can inform us not only about *what* but also *why* test persons may find elements relevant.

5. ACKNOWLEDGMENTS

We wish to thank the Danish HCI Forum for a stimulating discussion that led to the idea for this paper, and Gabriella Kazai for fruitful debates about ideal relevance assessments.

6. REFERENCES

1. Beaulieu, M. (1997): Experiments on interfaces to support query expansion. *Journal of Documentation*, 53(1), p. 8-19. (Special issue on Okapi)
2. Borlund, P. (2000): *Evaluation of interactive information retrieval systems*. Åbo: Åbo Akademi University Press, 276 p. (PhD dissertation)
3. Fuhr, N., Lalmas, M. and Malik, S. (2005): *Advances in XML Information Retrieval: Third International Workshop of the Initiative for the Evaluation of XML Retrieval, INEX 2004, Dagstuhl Castle, Germany, December 6-8, 2004, Revised Selected Papers*. Berlin: Springer. (LNCS ; 3493)
4. Hansen, J. Paulin. (1991): The use of eye mark recordings to support verbal retrospection in software testing. *Acta Pcychologica*. 76, pp. 31 - 49
5. Järvelin, K. and Kekäläinen, J. (2002): Cumulated gain-based evaluation of IR techniques. *ACM Transactions on Information Systems*, 20(4), 422-446.
6. Kazai, G., Lalmas, M. and de Vries, A. P. (2004): The overlap problem in content-oriented XML retrieval evaluation. In: *Proceedings of SIGIR 2004*, p. 72-79.
7. Kekäläinen, J. and Järvelin, K. (2002): Using graded relevance assessments in IR evaluation. *Journal of the American Society for Information Science and Technology*, 53(13), 1120-1129.
8. Pharo, N. and Nordlie, R. (2005): Context matters – an analysis of assessments of XML documents. In: *Proceedings of CoLIS5*, p. 238 – 248. (LNCS 3507)
9. Sormunen, E. (2002): Liberal relevance criteria of TREC : counting on negligible documents? In: *Proceedings of SIGIR 2002*, p. 324-330.
10. Tombros, A., Larsen, B. and Malik, S. (2005): The Interactive Track at INEX 2004. In: Fuhr, N., Lalmas, M. and Malik, S. eds. *Advances in XML Information Retrieval: Third International Workshop of the Initiative for the Evaluation of XML Retrieval, INEX 2004, Dagstuhl Castle, Germany, December 6-8, 2004, Revised Selected Papers.* Berlin: Springer, p. 410-423. (Lecture Notes in Computer Science ; 3493)
11. Toms, E. G., O'Brian, H. L., Kopak, R. & and Freund, L. (2005): Searching for relevance in the relevance of search. In: *Proceedings of CoLIS5*, p. 59 – 78. (LNCS 3507)
12. White, R.W., Ruthven, I., Jose, J.M. (2002). Finding relevant documents using top ranking sentences: an evaluation of two alternative schemes. In: *Proceedings of SIGIR 2002*, pp. 57-64.

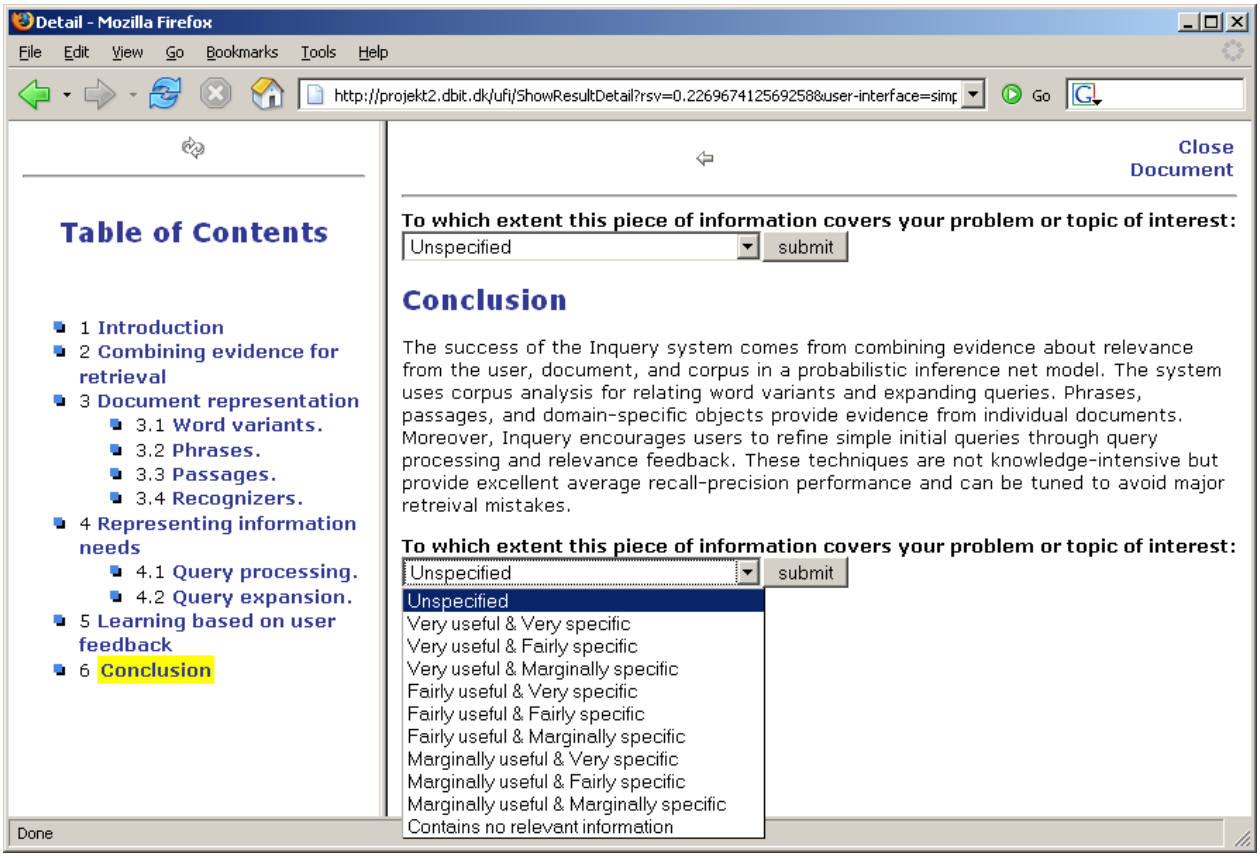

Figure 1. The HyREX XML IR system with prototype interface as used in the INEX 2004 interactive track [see 10]. Detailed component view containing the full text of the component and a table of contents for the whole document, and showing the relevance assessment scale.