Understanding Differences between Search Requests in XML Element Retrieval

Jaap Kamps[1,2] Birger Larsen[3]

[1] Archives and Information Studies, University of Amsterdam, Amsterdam, The Netherlands
[2] Informatics Institute, University of Amsterdam, Amsterdam, The Netherlands
[3] Information Studies, Royal School of Library and Information Science, Copenhagen, Denmark

ABSTRACT

XML retrieval, a very active branch of IR, studies the focused retrieval of semi-structured data. Although much progress has been made, especially through the annual INitiative for the Evaluation of XML retrieval (INEX), very little is known about XML element retrieval in action: What do users expect from an element retrieval system? What kind of information needs do they have? What sort of results do they request? Etc. In an effort to recover some of the answers, an extensive questionnaire was part of the peer topic creation process at INEX 2006. In this paper we present an analysis of the responses of topic authors. Our main general finding is that there is a great variety in the responses, and hence in the expectations about XML element retrieval.

Categories and Subject Descriptors

H.2 [**Database Management**]: H.2.3 Languages—*Query Languages*; H.3 [**Information Storage and Retrieval**]: H.3.1 Content Analysis and Indexing; H.3.3 Information Search and Retrieval; H.3.4 Systems and Software; H.3.7 Digital Libraries

General Terms

Measurement, Performance, Experimentation

Keywords

XML Retrieval, Search Requests, User Expectations

1. INTRODUCTION

Research in XML element retrieval attempts to take advantage of the structure of explicitly marked up documents to provide more focused retrieval results [7]. A special problem for this research area is that we have little knowledge about the expectations that potential users might have: As research in XML element retrieval is in its initial stages there are no operational systems with established user groups from which such expectations can be learned [15]. In this paper, we study a particular group of users who have worked intensively with an XML element retrieval system, in order to get some idea of their expectations of such systems.

SIGIR 2006 Workshop on XML Element Retrieval Methodology
August 10, 2006, Seattle, Washington, USA
Copyright of this article remains with the authors.

The task of XML element retrieval is a much more complicated one than standard document retrieval. Not only must XML element retrieval systems be able to identify relevant content; in addition a suitable granularity of the returned elements must be decided on along with how to handle overlap among elements [9]. As a consequence the creation of test collections for XML element retrieval is a notable challenge in itself. The main research effort in this area has since 2002 been the INitiative for the Evaluation of XML Retrieval [INEX 7]. Mainly due to INEX, much progress has been made with dedicated retrieval techniques [e.g., 3, 6, 8].

In addition, INEX includes an interactive track from 2004 onwards that has as purpose to investigate the behavior of users as they interact with XML element retrieval systems [10, 13]. However, the users studied in the INEX interactive track have no prior experience in searching XML element retrieval systems and only interact with them in a single session. Therefore the track is to a certain extent limited to studying novice users.

A hitherto unstudied user group is the authors of topics for the test collection. The test collection topics are created collectively by members of the research groups participating in INEX. The topics are created through a number of steps which involve repeated exploratory searches in an XML element retrieval system, and the assessment of a large number of elements [11]. Thus on the one hand the task is a very specific one, but on the other hand it demands that the system is used extensively over several days. The topic authors thereby become one of the most experienced groups of XML element retrieval users. Because of the collaborative effort most users in this group are drawn from people close to the participating research groups and are as a result more closely resembling real users and real tasks than in most other IR research settings [e.g., TREC 14].

Therefore we added an on-line topic questionnaire in INEX 2006, which the topic authors completed immediately after submitting the final version of their topics. The questionnaire consisted of 19 questions about the topic familiarity, the type of information requested and expected, results presentation, and the use of structured queries. It is important to stress that the questionnaire data is collected in the initial phase of the INEX campaign, before the retrieval tasks, metrics, or assessment instructions have been finalized. From the responses we hope to learn, in an indirect manner, more about user expectations for XML element retrieval systems. Moreover, we plan to distribute these data together with the test collection and hope that they will prove to be a valuable addition.

Table 1: Number of candidate topics per topic author at INEX 2006.

Min	Max	Median	Mean	Std. deviation
1	6	2	2.41	1.69

Table 2: (B1) How familiar are you with the subject matter of the topic?

Answer	Frequency	Percentage
Not familiar	8	4%
	139	71%
Very familiar	48	25%

Table 3: (B2) Would you search for this topic in real-life?

Answer	Frequency	Percentage
yes	186	95%
no	9	5%

Table 4: (B3) Does your query differ from what you would type in a web search engine?

Answer	Frequency	Percentage
yes	33	17%
no	162	83%

The paper is structured as follows: Section 2 presents the questionnaire and presents an analysis of the main results. Section 3 discusses relations between the questions, and Section 4 gives conclusions and points to future work.

2. CANDIDATE TOPIC QUESTIONNAIRE

An IR test collection consists of a collection of documents, a set of search topics, and relevance judgments. For INEX 2006, the document collection is an XML'ified version of the English Wikipedia [5]. At INEX, search requests or topics are authored (and also judged) by the INEX participants [11].

At INEX 2006, 81 different topic authors submitted a total of 195 topics (see Table 1 for some statistics). A total of 125 of the candidate topics have been selected as the topics for the INEX 2006 ad hoc retrieval tasks.

Directly after submitting a candidate topic (see [11] for details), the topic author was presented with a new page containing a questionnaire consisting of 19 questions and an open space for comments on the questionnaire. The 19 questions dealt with various issues related to the background of the search request and the topic author.

- the topic author's familiarity with the topic;
- the type of information requested;
- the type of search results expected;
- the type of results presentation preferred; and
- the meaning of structured queries.

Below we summarize the responses to all 19 questions of the candidate topic questionnaire at INEX 2006.

2.1 Topic Familiarity

The topic questionnaire featured three questions dealing with the familiarity and naturalness of the topics:

B1 How familiar are you with the subject matter of the topic? (yes/no)

B2 Would you search for this topic in real-life? (yes/no)

B3 Does your query differ from what you would type in a web search engine? (yes/no)

Table 2 shows the familiarity with the subject matter of the topic at hand.[1] It is reassuring that the vast majority of topic authors is familiar with the subject, although there are still 4% of the topics where topic authors venture into unfamiliar terrain.

Table 3 shows whether the topic corresponds to a the real-life search. The responses are overwhelmingly yes. For topic authors answering no, there was a follow-up question asking for their motivation. Typical responses where knowing the answer already, or not being interested in the answer.

At INEX 2006, the topic statement consists of a short keyword title, and an optional structured query [11]. Table 4 shows whether the provided topic statement differs from what the topic author would issue as a query to a web search engine. For 83% of the topics, there is no difference. For topic authors answering yes, there was a follow-up question asking for their motivation. For many of the topics that are different, the topic authors consider the structured query as the search request (and mention that this is not supported on standard web search engines).

Based on the three questions, we can conclude that the majority of topic authors search for familiar subject matter, provide a real-life search task, and provide a standard web search engine query.

2.2 Type of Information Requested

The questionnaire contains seven questions dealing with the type of information requested:

B4 Are you looking for very specific information? (yes/no)

B5 Are you interested in reading a lot of relevant information on the topic? (yes/no)

B6 Could the topic be satisfied by combining the information in different (parts of) documents?

B7 Is the topic based on a seen relevant (part of a) document? (yes/no)

B8 Can information of equal relevance to the topic be found in several documents? (yes/no/don't know)

B9 Approximately how many articles in the whole collection do you expect to contain relevant information?

B10 Approximately how many relevant document parts do you expect in the whole collection?

[1]Due to a problem with the form for question B1, categories 2 and 4 of the original five point scale have been collapsed in the answers logs. We derive a three point scale for B1 by grouping the answers to categories 2, 3, and 4 as a single intermediate category.

Table 5: (B4) Are you looking for very specific information?

Answer	Frequency	Percentage
yes	114	58%
no	81	42%

Table 6: (B5) Are you interested in reading a lot of relevant information on the topic?

Answer	Frequency	Percentage
yes	123	63%
no	72	37%

Table 7: (B6) Could the topic be satisfied by combining the information in different (parts of) documents?

Answer	Frequency	Percentage
yes	160	82%
no	35	18%

Table 8: (B7) Is the topic based on a seen relevant (part of a) document?

Answer	Frequency	Percentage
yes	74	38%
no	121	62%

Table 9: (B8) Can information of equal relevance to the topic be found in several documents?

Answer	Frequency	Percentage
yes	163	84%
no	12	6%
don't know	20	10%

Table 5 shows whether topics are asking for very specific information. For 58% of the topics, the response is yes, indicating many topics can likely be answered by a relatively small amount of text.

Table 6 shows whether topics authors are interested in reading a lot of relevant information. Now, for 63% of the topics the answer is yes, indicating that recall is appreciated for most of the topics.

Table 7 shows whether the topics can be satisfied by combining information in different (parts of) documents. Here, for no less than 82% of the topics, the anwers is yes. This can be interpreted to indicate that many topics are multi-faceted.

These three questions, B4-6, try to assess the scope of the topics. The outcome is mixed: B4 indicates a narrow scope, but B6 indicates a broad scope. We return to the relation between the responses to these questions in Section 3 below.

Table 8 shows whether topics are based on a seen relevant document. Here, for 62% of the topics, the response is no, indicating that these are clearly not "known-item" topics.

Table 9 shows whether information of equal relevance can be found in different documents. For 84% of the topics, the reponse is yes, indicating that these are informational search topics rather than navigational topics [1].

Tables 10 and 11 show some statistics on the expected number of articles and elements with relevance. The distributions are both fairly skewed, but showing that relevance is expected in a wide range of articles and elements.

Table 10: (B9) Approximately how many articles in the whole collection do you expect to contain relevant information?

Min	Max	Median	Mean	Std. deviation
2	15,000	20	128	1097

Table 11: (B10) Approximately how many relevant document parts do you expect in the whole collection?

Min	Max	Median	Mean	Std. deviation
2	20,000	50	289	1671

Table 12: (B11) Could a relevant result be (check all that apply)?

Answer	Frequency	Percentage
a single sentence	81	42%
a single paragraph	139	71%
a single (sub)section	170	87%
a whole article	160	82%

Table 13: (B12) Can the topic be completely satisfied by a single relevant result?

Answer	Frequency	Percentage
yes	74	38%
no	121	62%

These four questions, B7-10, try to assess to what extent search requests resemble known-item search topics or ad hoc retrieval topics. Based on the responses, we can conclude that the topics are predominantly general informational topics.

2.3 Type of Results Expected

The questionnaire has four questions zooming in on the type of search results expected:

B11 Could a relevant result be (check all that apply): a single sentence; a single paragraph; a single (sub)section; a whole article.

B12 Can the topic be completely satisfied by a single relevant result? (yes/no)

B13 Is there additional value in reading several relevant results? (yes/no)

B14 Is there additional value in knowing all relevant results? (yes/no)

Table 12 shows the expected result granularity (note that multiples answers are possible). Some observations present themselves. First, for no less than 42% of the topics a single sentence could be a relevant result, indicating a very specific information need that can be answered by a single sentence. Second, for no less than 82% of the topics a whole article could be a relevant result.

Table 13 shows which topics can be completely satisfied by a single relevant result. For 38% of the topics this is the case.

Table 14: (B13) Is there additional value in reading several relevant results?

Answer		Frequency	Percentage
Not important	1	7	4%
	2	15	8%
	3	36	18%
	4	71	36%
Very important	5	66	34%

Table 15: (B14) Is there additional value in knowing all relevant results?

Answer		Frequency	Percentage
Not important	1	21	11%
	2	41	21%
	3	49	25%
	4	53	27%
Very important	5	31	16%

Table 16: (B15) Would you prefer seeing?

Answer	Frequency	Percentage
only the best results	82	42%
all relevant results	106	54%
don't know	7	4%

Table 17: (B16) Would you prefer seeing?

Answer	Frequency	Percentage
isolated document parts	69	35%
the article's context	105	54%
don't know	21	11%

Table 14 shows the importance of reading several relevant results. For 70% of the topics there is clear importance (4 or 5 on the 5-point scale).

Table 15 shows the importance of reading all relevant results. Now we see a very even distribution of topics over importance.

These three questions, B12-14, try to assess the relative importance of precision and recall for the search requests. We see that for most topics, the topic authors are interested in recall.

2.4 Results Presentation

The questionnaire has two questions zooming in on result presentation:

B15 Would you prefer seeing: only the best results; all relevant results; don't know

B16 Would you prefer seeing: isolated document parts; the article's context; don't know

Table 16 shows how many of the relevant results topic authors prefer to see. The outcome is mixed: for 54% of the topics, all results should be shown, and for 42% of the topics only the best results need to be shown.

Table 17 shows whether results should be shown in their original article's context. For 54% of the topics, a presentation in context is preferred, whereas for 35% of the topics isolated results are preferred.

Table 18: (B17) Do you assume perfect knowledge of the DTD?

Answer	Frequency	Percentage
yes	24	12%
no	171	88%

Table 19: (B18) Do you assume that the structure of at least one relevant result is known?

Answer	Frequency	Percentage
yes	65	33%
no	130	67%

Table 20: (B19) Do you assume that references to the document structure are vague and imprecise?

Answer	Frequency	Percentage
yes	121	62%
no	74	38%

These two questions, B15-16, show that topic authors have different preferences on the presentation of XML element retrieval results.

2.5 Structured Queries

The questionnaire featured three questions dealing with structured queries, the so-called content-and-structure (CAS) queries formulated in the NEXI language [16].

B17 Do you assume perfect knowledge of the DTD? (yes/no)

B18 Do you assume that the structure of at least one relevant result is known? (yes/no)

B19 Do you assume that references to the document structure are vague and imprecise? (yes/no)

Even though these questions where also optional (because formulating a structured query was no requirement), the questions were answered for all topics.

Table 18 shows whether topic authors assumed a perfect knowledge of the collection's mark-up structure. As it turned out, for 12% of the topics, perfect knowledge of the DTD is assumed.

Table 19 shows whether the mark-up structure of at least one relevant result is known. Now, for 33% of the topics it is assumed that the structure at least one result is known.

Table 20 shows how to interpret structural references in the search request. Here, for 62% of the topics, structural references are considered vague and imprecise. However, in 38% of the topics, the structural hints are meant to be interpreted literally.

These three questions, B17-19, address the meaning of structured queries in XML element retrieval. The results show that for a majority of topics the structural references are merely search hints, but that for a seizable fraction structure should be taken seriously.

3. RELATIONS

In this section, we analyze the relation between responses to different questions in the questionnaire. Table 21 show the relations between pairs of questions in the questionnaire. We will discuss these relations in detail.

First we focus on topic familiarity.

Table 21: Relationship between answers for pairs of questions (chi-square test at percentiles 0.95 and 0.99).

	B1	B2	B3	B4	B5	B6	B7	B8	B12	B13	B14	B15	B16	B17	B18	B19
B1																
B2	-															
B3	-	-														
B4	0.99	-	-													
B5	-	-	0.95	0.99												
B6	-	0.99	-	0.95	-											
B7	-	-	-	0.95	-	-										
B8	-	-	-	-	-	0.95	-									
B12	-	-	-	0.99	0.99	-	-	-								
B13	-	-	-	-	0.99	0.99	-	0.95	0.99							
B14	0.95	-	-	-	0.99	-	-	-	0.99	0.99						
B15	-	-	-	-	0.99	0.95	-	-	0.99	0.99	0.99					
B16	-	-	-	-	0.95	-	0.99	-	-	-	-	-				
B17	0.95	-	-	-	-	0.95	0.99	-	-	-	-	-	-			
B18	-	-	0.99	0.95	-	-	0.95	-	-	-	-	-	-	0.99		
B19	-	-	-	-	-	-	-	-	-	-	-	0.95	-	-	-	

B1,B4 Topics which the author is very familiar with are more often very specific.

B1,B14 Topics which the author is moderate familiar with, have a moderate importance of knowing all the relevant results.

B1,B17 Topics which the author is very familiar with the subject matter of the topic at hand, do more often assume perfect knowledge of the DTD.

Here, the relation between B1 and B4 is interesting: topic authors ask more specific queries about familiar subject matter. This simple observation has a bearing on the sort of users and tasks for which XML element retrieval system is most suitable.

Second, we focus on the naturalness of the topic and query.

B2,B6 Real-life topic are more often satisfied by combining information in different (parts of) documents.

B3,B5 Topic statements identical to Web search engine queries, make reading a lot of information more interesting.

B3,B18 Topic statements identical to Web search engine queries, less often assume that the structure of at least one relevant result is known.

These relations suggest that these topics are resonating closely with the sort of search request issued in real world information gathering.

Third, we look at the specificness of the topics:

B4,B5 Topics asking for very specific information, make reading a lot of information less interesting.

B4,B6 Topics asking for very specific information, do not expect answers from combining information in different (parts of) documents.

B4,B7,B18 Topics asking for for very specific information, are often based on a seen relevant document; and more often assume that the structure of at least one relevant document is known.

B4,B12 Topics asking for very specific information, are more often completely satisfied with a single relevant result.

Here the relation between B4 and B6 is clearly inverse, indicating that very specific topics are mono-faceted. The general suggestion is that specific topics form a category with distinct characteristics.

Fourth, we discuss the importance of reading relevant information:

B5,B12,B13,B14,B15 Topics for which it is of interest to read a lot of relevant information; are less often completely satisfied with a single relevant result; make reading several relevant results more important; make knowing all results more important; and make seeing all results more important.

B5,B16 Topics for which it is of interest to read a lot of relevant information, it is preferred to read information in the article's context.

Here the general suggestion is that topics for which reading a lot of relevant information is important also from a distinct category, which, considering the inverse relation between B4 and B5, is roughly complementing the category of specific topics.

Fifth, we focus on topics that can be satisfied by the combination of information in different (parts of) documents:

B6,B8,B13,B15 Topics that can be satisfied by combining information in different (parts of) documents, more often have information of equal relevance in different documents; and make reading several relevant results more important; and make seeing all relevant results more important.

B6,B17 For all topics assuming perfect knowledge of the DTD, it is assumed that they can be satisfied by combining information in different (parts of) documents.

Given that B4 on topic specificity was inversely related with B6, these relations reaffirm the differences between specific topics, and topic for which reading a lot of relevant information is interesting.

Sixth, we continue with seen relevant documents:

B7,B16 For topics based on a seen relevant document, it is less often preferred to see the article's context.

B7,B17,B18 Topics based on a seen relevant document are more often assuming perfect knowledge of the DTD; and are more often assuming that the structure of at least one relevant document is known.

These relations clearly suggest the prior knowledge assumed on the part of the searcher for these topics.

Finally, seventh, we look at vague structural hints:

B15,B19 For topics with vague structural hints, it is more often preferred to see only the best results.

This is an interesting relation, which could be interpreted to mean to vague structural hints are provided to improve the ranking of certain elements.

The responses to the two numerical questions on the number of relevant articles and document parts, B9 and B10, are clearly related (Pearson correlation 0.9215), as may be expected.

We excluded above the responses to question B11 (about the granularity of potential relevant results) since multiple answers are possible. As it turns out, there are three relations between the responses to B11: sentence is related to paragraph, paragraph is related section, and section is related to article. For all other pairs of responses (e.g, sentence-article), we find no relation.

Finally, recall that most topic authors submitted multiple candidate topics. We analyzed the relation between the topic author and the questions above. For the twelve questions, B1, B3, B5, B7, B8, B11, and B14–B19, the responses are related to the particular topic author at hand. This suggests that some of the responses are mainly related to the topic at hand, whereas others are mainly related to the particular user.

4. CONCLUSIONS

Studying the expectations of the INEX topic authors as an example of an XML element retrieval user group has its advantages and disadvantages. The task they have performed is a highly specific and somewhat artificial one (that of producing a test collection topic) compared to the natural tasks of real users. However, the INEX topic authors are probably much closer to real users than in other IR test collection building efforts because they are mainly recruited among the participating research groups. It is therefore reassuring that most topic authors searched for familiar subject matter and real-life tasks using queries similar to web queries. In addition to this the Wikipedia collection covers a very broad range of subject matter, and the topic authors have generally extensive experience with XML element retrieval systems. Arguably, the results of this study will extend our understanding of what users expect from an XML element retrieval system.

Perhaps the most striking observation is that there is such great variety in the expectations of the topic authors. This may, in turn, indicate that there is a range of several different XML retrieval tasks types. This give broad support to the decision at INEX to define a number of distinct retrieval tasks [4]. In particular, we have found that there are a number of relations worth considering. Among these is that great topic familiarity lead to more specific topics, and that the specific topics tend to be mono-faceted and can be completely satisfied by a single relevant result. For more complex topics where moderate or even high recall is desired it is preferred to read more information and to present the results in the articles context. In addition, it appears that there are two distinct views on the meaning of structural hints: the majority regards them as only vague hints, but for a sizable fraction they should be taken seriously.

In this paper, we only reported the responses to the questionnaire as a survey amongst candidate topic authors, which can be construed as a particular group of XML element retrieval users. However, recall that 125 of the candidate topics were selected as the INEX 2006 ad hoc retrieval topics, and—at a later stage of the INEX campaign—the topic authors will be asked to make relevance judgments for pooled sets of elements. That is, the questionnaire data also becomes part of the evaluation test-suite that will be constructed during INEX 2006, providing valuable contextual data on the topics of request and their topic authors.

This enriched test set will have a number of unique features. First, it will allow to breakdown the set of topics in various meaningful categories, and zoom in on the relative performance for such a group of topics. Second, zooming on particular topic categories will help to explain diverging results between different techniques, tasks, and metrics. Third, it will reveal the importance of each of the variables measured in the questionnaire for the various INEX tasks [4]. Fourth, it may help us understand what are the fundamental differences between tasks, which will lead in turn to better retrieval techniques for individual tasks. In short, the rich contextual information from the topics questionnaire will significantly boost the value of the test suite constructed during INEX 2006, and greatly increase the potential reuse of the test suite in the future.

The Cranfield tradition of test collection development tries to abstract away from individual differences between assessors [17]. Yet at the same time, it is known for long that individual difference are one of the greatest sources of variation in relevance judgments and system failure [2, 12]. Given that the task of XML element retrieval is of a higher complexity that standard document retrieval, due to the document structure, the fine-grained judgments, and, perhaps, a lack of consensus on the precise retrieval task, it is more than plausible that individual differences have a much greater impact. The questionnaire data will shed light on the impact of these differences—even zoom in on the relative impact of specific features—and at the same time provide a handle to deal with them.

Acknowledgments

Thank you to all INEX 2006 topic authors, and to Mounia Lalmas and Saadia Maalik for their help with the questionnaire and data collection.

Jaap Kamps was supported by the Netherlands Organization for Scientific Research (NWO) under project numbers 612.066.513, 612.066.302, and 640.001.501.

Birger Larsen was supported by the NORSLIS Research School.

REFERENCES

[1] A. Broder. A taxonomy of web search. *SIGIR Forum*, 36(2):3–10, 2002.

[2] C. Buckley. Why current ir engines fail. In M. Sanderson, K. Järvelin, J. Allan, and P. Bruza, editors, *Proceedings of the 27th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 584–585. ACM Press, New York NY, USA, 2004.

[3] D. Carmel, Y. S. Maarek, M. Mandelbrod, Y. Mass, and A. Soffer. Searching XML documents via XML fragments. In C. Clarke, G. Cormack, J. Callan, D. Hawking, and A. Smeaton, editors, *Proceedings of the 26th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 151–158. ACM Press, New York NY, USA, 2003.

[4] C. Clarke, J. Kamps, and M. Lalmas. INEX 2006 retrieval task and result specification. In *INEX 2006*, 2006.

[5] L. Denoyer and P. Gallinari. The Wikipedia XML Corpus. *SIGIR Forum*, 40:64–69, 2006.

[6] N. Fuhr and K. Großjohann. XIRQL: A query language for information retrieval in XML documents. In D. H. Kraft, W. B. Croft, D. J. Harper, and J. Zobel, editors, *Proceedings of the 24th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 172–180. ACM Press, New York NY, USA, 2001.

[7] INEX. INitiative for the Evaluation of XML Retrieval, 2006. `http://inex.is.informatik.uni-duisburg.de/2006/`.

[8] J. Kamps, M. de Rijke, and B. Sigurbjörnsson. Length normalization in XML retrieval. In M. Sanderson, K. Järvelin, J. Allan, and P. Bruza, editors, *Proceedings of the 27th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 80–87. ACM Press, New York NY, USA, 2004.

[9] G. Kazai, M. Lalmas, and A. P. de Vries. The overlap problem in content-oriented XML retrieval evaluation. In *Proceedings of the 27th Annual International ACM SIGIR Conference*, pages 72–79. ACM Press, New York NY, USA, 2004.

[10] B. Larsen, S. Malik, and T. Tombros. The interactive track at INEX 2005. In N. Fuhr, M. Lalmas, S. Malik, and G. Kazai, editors, *Advances in XML Information Retrieval and Evaluation: Fourth Workshop of the INitiative for the Evaluation of XML Retrieval (INEX 2005)*, volume 3977. Springer Verlag, Heidelberg, 2006.

[11] B. Larsen and A. Trotman. INEX 2006 guidelines for topic development. In *INEX 2006*, 2006.

[12] T. Saracevic. Relevance: A review of and a framework for the thinking on the notion in information science. *JASIS*, 26:321–343, 1975.

[13] A. Tombros, B. Larsen, and S. Malik. The interactive track at INEX 2004. In *Advances in XML Information Retrieval. Third Workshop of the INitiative for the Evaluation of XML Retrieval, INEX 2004*, volume 3493 of *Lecture Notes in Computer Science*, pages 410–423. Springer Verlag, Heidelberg, 2005.

[14] TREC. Text REtrieval Conference, 2006. `http://trec.nist.gov/`.

[15] A. Trotman. Wanted: Element retrieval users. In A. Trotman, M. Lalmas, and N. Fuhr, editors, *Proceedings of the INEX 2005 Workshop on Element Retrieval Methodology*, pages 63–69. University of Otago, Dunedin New Zealand, 2005.

[16] A. Trotman and B. Sigurbjörnsson. Narrowed Extended XPath I (NEXI). In N. Fuhr, M. Lalmas, S. Malik, and Z. Szlávik, editors, *Proceedings of the Initiative for the Evaluation of XML Retrieval (INEX 2004)*, Lecture Notes in Computer Science, pages 16–40. Springer Verlag, Heidelberg, 2005.

[17] E. M. Voorhees. The philosophy of information retrieval evaluation. In C. Peters, M. Braschler, J. Gonzalo, and M. Kluck, editors, *Evaluation of Cross-Language Information Retrieval Systems, CLEF 2001*, volume 2406 of *Lecture Notes in Computer Science*, pages 355–370. Springer, 2002.