Context matters: an analysis of assessments of XML documents

Nils Pharo and Ragnar Nordlie

Oslo University College, Faculty of Journalism, Library and Information Science,
Postboks 4 St. Olavs plass, N-0130 Oslo, Norway
{nils.pharo, ragnar.nordlie}@jbi.hio.no

Abstract: The paper analyses searchers' assessments of usefulness and specificity on different levels of granularity in XML-coded documents. Documents are assessed on 10 usefulness/specificity combinations and on the granularity levels of article, section, and subsection. Overlapping judgements show a remarkable lack of consistency between searchers. There is an inverse relationship between articles and sections both in the assessment of specificity and of usefulness, indicating that retrieval on different granularity levels are a useful feature of a retrieval system. Searchers find the full article more useful when they assess the same document both on the article and section level indicating that there is a need to provide context to the sections and subsections when presenting result list of XML-documents.

Published as: Pharo, N. & Nordlie, R. (2005). Context Matters: An Analysis of Assessments of XML Documents. In: F. Crestani and I. Ruthven (Eds.): CoLIS 2005, LNCS 3507 (pp. 238 – 248). Berlin Heidelberg: Springer-Verlag.

1 Introduction

The eXtensible Markup Language (XML) is increasingly becoming the standard for content representation on the Web. In this paper the focus is on XML used for representing semi-structured documents, i.e., documents with a certain amount of systematically occurring elements mixed with longer bits of unstructured full text. Scientific articles represent good examples of semi-structured documents, where the content partly consist of specific formally defined elements such as titles, captions, footnotes, headings, formulas etc, as well as elements representing unstructured sections of full text such as abstracts, subsections, paragraphs etc. These elements are for a large part used in order to serve publishing and presentation purposes, but to exploit these structural elements in information retrieval is an appealing idea, e.g., by developing ranking algorithms that combine element names and content.

One of the presumed advantages of XML-based information retrieval is that the XML coding will enable retrieval systems to present searchers with search results consisting of the document elements presumed to be most relevant to their problem [1]. The underlying assumption is that searchers should retrieve as much, but not more of the document than is necessary to satisfy their information need. We wish to investigate the validity of this assumption. In this paper we present a study of searchers' relevance assessments of different levels of granularity in XML documents. Our main research question has been to investigate how different levels of granularity influence searchers' evaluation and their ability to evaluate. In this study the lowest level of granularity is sections and subsections of articles.

2 Previous work

Both outside of and particularly within the framework of the INEX family of experiments, much has been written on various aspects of information retrieval in structured documents, and a particular focus within INEX has been on metrics for retrieval evaluation in such settings, see for instance [5], but there are few investigations of searcher behaviour in this connection. [2] analyze which parts of structured documents searchers access (in their case a structured collection of Shakespeare texts), but their focus is on task performance and interface design, not on

relevance assessments. A brief summary of findings from the INEX interactive track is presented in [9]. Our investigation elaborates some of the general findings referenced here. There are a number of investigations which discuss the problems connected with such aspects of user assessments as for instance the use of graded relevance assessments, e.g. [3], and an extensive literature on the problems of consistency in relevance judgements, see for instance [8].

3 Method

At present the largest set of available data on how searchers evaluate XML documents on different levels of granularity stems from the international Initiative for the Evaluation of XML retrieval (INEX). We chose to use data collected from this initiative, thus limiting our ability to control factors such as participants and tasks. In this chapter, we first describe the INEX initiative, which is followed by a part presenting how we analysed the data.

3.1 The INEX initiative

INEX was established in 2002 in order to provide "an infrastructure to evaluate the effectiveness of content-oriented XML retrieval systems" [4]. INEX builds its experimental design on the TREC model, with a test collection which consists of topics/tasks (submitted by the participating groups), documents (approximately 12 000 articles from a selection of IEEE Computer society's journals) and relevance assessments provided by the participants, thus making it possible to compute the retrieval effectiveness of different matching algorithms.

A new interactive track was introduced in INEX in 2004 [9] which aimed at focusing on how searchers performed when solving the tasks (which for this experiment were formulated following Borlund's [1] simulated work task procedure). The INEX 2004 Interactive Track (http://inex.is.informatik.uni-duisburg.de:2004/tracks/int/) is a collective effort by ten different research groups at sites in Asia, Australia and Europe. The data are collected at the different sites from searchers who were each given two search tasks of different complexity and performed searches following precise guidelines from the track organizers:

- The Hyrex experimental IR system was used with a specific interface developed for the INEX interactive track [7]
- Searchers were allowed to spend a maximum of 30 minutes working on each task
- Searchers were requested to assess all document elements they chose to view on a ten-point relevance scale (see Table 1 for the relevance scale)

Table 1. Relevance scale

Grade	Description
A	Very useful & Very specific
B	Very useful & Fairly specific
C	Very useful & Marginally specific
D	Fairly useful & Very specific
E	Fairly useful & Fairly specific
F	Fairly useful & Marginally specific
G	Marginally useful & Very specific
H	Marginally useful & Fairly specific
I	Marginally useful & Marginally specific

J	Contains no relevant information
U	Unspecified

The system is designed with a simple search interface where searchers can input queries to the system. The result list contains four different granularity levels of documents: whole articles, sections, subsection level 1, and subsection level 2. When selecting a (part of an) article the searcher also is presented with a table of contents to the other parts (sections/subsections) of the article. The system provides searchers with the opportunity to assess the relevance level of the different entries in the result list. The relevance levels are based on two dimensions of relevance, "usefulness" and "specificity". Usefulness has to do with the exhaustiveness of the documents' treatment of the question topic, in fact in the other tracks at INEX 04 "exhaustiveness" has been used to signify this dimension rather than usefulness. Specificity deals with the extent to which the retrieved article (part) is focussed on the topic of the searcher's task. The ten-point relevance scale combined three different levels (from "marginally" via "fairly" to "very") of specificity and exhaustiveness in addition to the option of judging the document (part) non-relevant.

There were ten research institutions around the world participating in the study, each site was required to collect data from at least eight volunteers. The data were collected following the guidelines from the INEX Interactive Track organisers: participants were first given a brief introduction to the experiment and the Hyrex system, before and after the experiment they were asked to fill out general questionnaires, the searchers selected one task from each of two task categories, before and after each task they were asked to answer task-related questionnaires. The search tasks were formulated as simulated work task situations [1], meaning that the tasks were also placed in a more specific context, giving the searchers more information about why the information is needed.

Two tasks belonged to the Background category (*B*), the other two to the Comparison category (*C*). Table 2 contains the four tasks as they were presented to the searchers.

All transaction in the systems are logged in XML and plain text format, including queries, viewed document element, search paths, assessments, time spent etc.

We have analysed transaction logs from the search sessions in order to look at the distribution of different levels of relevance assessments at the various document levels.

Table 2. Simulated work tasks in INEX Interactive track 2004

Task ID: B1 You are writing a large article discussing virtual reality (VR) applications and you need to discuss their negative side effects. What you want to know is the symptoms associated with cybersickness, the amount of users who get them, and the VR situations where they occur. You are not interested in the use of VR in therapeutic treatments unless they discuss VR side effects.	**Task ID: B2** You have tried to buy & download electronic books (ebooks) just to discover that problems arise when you use the ebooks on different PC's, or when you want to copy the ebooks to Personal Digital Assistants. The worst disturbance factor is that the content is not accessible after a few tries, because an invisible counter reaches a maximum number of attempts. As ebooks exist in various formats and with different copy protection schemes, you would like to find articles, or parts of articles, which discuss various proprietary and covert methods of protection. You would also be interested in articles, or parts of articles, with a special focus on various disturbance factors surrounding ebook copyrights.

Task ID: C1 You have been asked to make your Fortran compiler compatible with Fortran 90, and so you are interested in the features Fortran 90 added to the Fortran standard before it. You would like to know about compilers, especially compilers whose source code might be available. Discussion of people's experience with these features when they were new to them is also of interest.	**Task ID: C2** You are working on a project to develop a next generation version of a software system. You are trying to decide on the benefits and problems of implementation in a number of programming languages, but particularly Java and Python. You would like a good comparison of these for application development. You would like to see comparisons of Python and Java for developing large applications. You want to see articles, or parts of articles, that discuss the positive and negative aspects of the languages. Things that discuss either language with respect to application development may be also partially useful to you. Ideally, you would be looking for items that are discussing both efficiency of development and efficiency of execution time for applications.

3.2 Data analysis

In our study we have used the transaction logs from nine sites, in all 140 sessions. The sessions contained 1835 relevance assessments, out of which 1259 were between A and I, i.e. the article element was considered relevant to some degree. We have only made limited use of the data collected from the various questionnaires since our aim has been to look at the general distribution of relevance assessments over article elements rather than taking into account individual factors affecting the assessments. We are, however, aware that factors such as search experience and task knowledge influence the choices of individual searchers.

Table 3. Excerpts from log file

SearcherID	ArticleID	Article element	Grade
cmpinfscnor_searcher002_C	/cs/1998/c2039	/article[1]/bdy[1]/sec[1]	B
cmpinfscnor_searcher002_C	/cs/1998/c2039	/article[1]/bdy[1]/sec[3] /ss1[2]	I
cmpinfscnor_searcher002_C	/cs/1998/c2039	/article[1]/bdy[1]/sec[3] /ss1[3]	E

In Table 3 we see that a searcher has assessed three different parts of one article, section 1 (sec[1]), and subsections 2 and 3 in section 3.

In order to investigate our research problem we have investigated the following:

1. the distribution of assessments over article elements, independent of individual searchers, this provides information on what granularity level the searchers generally performed relevance assessments
2. the relationship between an individual searcher’s assessments of different elements of the same article
3. the distribution of all assessments for one specific article, which provides information about assessment consistency

4 Findings

4.1 Relationship between granularity and assessments

In total, searchers assessed slightly less than 600 individual documents, of which about 15% were full articles and 85% were sections or subsections of articles (coded with XML codes sec, ss1 or ss2). We do not, unfortunately, know the total distribution of sections and subsections in the 12 000 articles in the database thus we do not know if this reflects the general distribution. Of the 1835 assessments made by the searchers, 24% were article assessments and 76% were assessments of section or subsections. This means that searchers showed a marked tendency towards preferring to assess articles over sections of articles. Of the 1835 assessments, slightly less than 30% were "J", indicating no relevant information, and a small proportion were judged "unspecified". We have chosen to disregard these negative assessments in our further investigation of the material. It is difficult to judge from the logs why searchers have just chosen not to judge some of the documents they find unusable while they give others a negative assessment, so we feel this figure is burdened with too much uncertainty. This leaves us with 1259 individual assessments, distributed as shown in table 4.

Table 4. Distribution of assessments over document elements

	Article	Section	SS1	SS2	
A	33 (13.5%)	164 (67.2%)	42 (17.2%)	5 (2.1%)	244 (100%)
B	32 (28.1%)	56 (49.1%)	24 (21.0%)	2 (1.8%)	114 (100%)
C	9 (31.0%)	15 (51.7%)	5 (17.2%)	-	29 (99.9%)
D	28 (21.0%)	66 (49.6%)	35 (26.3%)	4 (3.0%)	133 (100%)
E	40 (25.0%)	76 (47.5%)	42 (26.3%)	2 (1.2%)	160 (100%)
F	43 (41.7%)	44 (42.7%)	12 (11.7%)	4 (3.9%)	103 (100%)
G	9 (14.3%)	41 (65.1%)	11 (17.5%)	2 (3.2%)	63 (100.1%)
H	23 (18.3%)	78 (61.9%)	21 (16.7%)	4 (3.2%)	126 (100.1%)
I	89 (31.0%)	125 (43.3%)	62 (21.6%)	11 (3.8%)	286 (100%)
Total	306	665	254	34	1259

From table 4, it appears that the distribution of ss1 and ss2 elements deviates little from the average for any of the categories. The most significant deviation from the normal is the relatively low proportion of "A" judgements on the article level and the comparably high proportion of "A"s on the section level. Since "A" includes both maximum specificity and usefulness, and a section of an article might be expected to treat a topic with more specificity than would the entire article, this is no surprise. Tables 5 and 6 show the relative influence of the two relevance dimensions.

Table 5. Distribution of various levels of specificity over document elements

	Article	Section	SS1	SS2	
Highly spec. A-D-G	70 22.9%	271 40.8%	88 34.6%	11 32.4%	440
Fairly spec. B-E-H	95 31.0%	210 31.6%	87 34.3%	8 23.5%	400
Marginally spec. C-F-I	141 46.1%	184 27.7%	79 31.1%	15 44.1%	419
Total	306 100%	665 100.1%	254 100%	34 100.1%	1259

Table 6. Distribution of various levels of usefulness over document elements

	Article	Section	SS1	SS2	
Highly useful A-B-C	74 24.2%	235 35.3%	71 28.0%	7 20.6%	387
Fairly useful D-E-F	111 36.3%	186 28.0%	89 35.0%	10 29.4%	396
Marginally useful G-H-I	121 39.5%	244 36.7%	94 37.0%	17 50.0%	476
Total	306 100%	665 100%	254 100%	34 100%	1259

As expected, there is a clear inverse relationship between articles and sections in the assessment of specificity; it is apparently (and intuitively) easier to relate the notion of specificity to the section level than to an entire article. It is more difficult to explain the somewhat slighter but still inverse relationship between articles and sections when it comes to judging usefulness. This is, for instance, in opposition to the INEX experiment designers' rules for assessing XML-coded parts of documents, which state that no sub-element can have a lower degree of exhaustivity than the mother element. One would intuitively think that if a section of an article is useful as the answer to a query, the entire article will be useful as well. It is possible that the term "usefulness" is difficult for the searchers to relate to in a setting where the problem which is the basis for judging the material is imposed on them rather than taken from their real-life situation. It might possibly have been easier for them if the searchers were asked to judge "exhaustivity" instead, which is the case for the non-interactive tracks in INEX. It may also be that the combined relevance dimensions makes it difficult to distinguish between "specificity" and "usefulness" – table 4 shows clearly that the three grades which give equal weight to the two measures (A, E and I for high/high, fairly/fairly and marginally/marginally, respectively) are much more heavily used than the others. Again, the use of two separate measures might have provided a more realistic representation of searcher assessments.

4.2 Assessment overlap

Tables 4-6 show the distribution of assessments without regard to individual searchers or individual search sessions. We find, in general, an increase in both usefulness and in specificity when searchers deal with smaller article element than when they address (and assess) the article as a whole. To investigate whether this is also the case when individual searchers have the chance to see and judge both the full article and its separate sections, we identified the assessment of all overlapping article elements, or *elements*, for each session, i.e. we identified each occurrence in the transaction log where one element and one or more of its sub-elements are assessed in the same session. In total there were 143 such assessments.

In order to identify increase and decrease in assessed usefulness and specificity we treated the two dimensions of the relevance grades separately. Grades A, B, and C were given the score 3 for usefulness; D, E, and F score 2; and G, H, and I scored 1. For specificity grades A, D, and G scored 3; B, E, and H scored 2; whereas C, F, and I were given the score 1. Now we could treat each assessment separately with respect to usefulness and specificity, and thus identify increase and decrease of assessed relevance for overlapping elements. An example is shown in the excerpts in Table 7.

Table 7. Excerpts from log file with overlapping assessments

SearcherID	ArticleID	Article element	Grade
dbdk_searcher012_B	/co/1995/r6057	/article[1]	F
dbdk_searcher012_B	co/1995/r6057	/article[1]/bdy[1]/sec[5]	A

In the Table 7 example we see that the searcher has assessed the article with an "F", meaning it is fairly useful (score: 2) and marginally specific (score: 1). Section 5, however, the searcher thinks is both very useful (score: 3) *and* very specific (score: 3). In this example we see that both the assessed usefulness and specificity increases with the increased document granularity.

The results of a similar treatment of all overlapping assessments in the transaction logs are presented in Table 8. The table should be read from the perspective of the assessment of the super-element, so that increase or decrease is from super-element assessment (e.g. article) to sub-element (e.g. section).

Table 8. Relevance assessment change in overlapping article elements

	Usefulness		Specificity	
Increase	17	(12 %)	36	(25 %)
Unchanged	66	(46 %)	68	(48 %)
Decrease	60	(42 %)	39	(27 %)

From Table 8 we can see that in almost half of all cases there is neither a decrease nor an increase in usefulness or specificity when lower-level elements are assessed. This means that the searcher most often find no difference in relevance between sub-elements and super-element.

The table also shows that searchers are much more likely to assess the sub-elements as less useful than the opposite. This indicates that searchers find the broader article or section more "useful" than the smaller sections and subsections as sources of information. This stands in apparent opposition to the findings in table 6, where we find proportionally more sections than articles judged "highly useful". As mentioned above, what is meant by "useful" is not clearly defined, and this is a source of uncertainty in any attempt to explain the discrepancy. It may seem, however, that even if judged independently a section seems more useful than a full article, the article in its entirety when seen in connection with the sections still offer more towards the searches' problem resolution. A better explanation might be found if we had been able

to consider the sequence of the assessments to see if the level of usefulness were influenced by the order in which the searcher viewed the article and the sections.

Table 8 also reveals that there is no clear tendency with respect to increase or decrease in assessed specificity of sub-elements. One might expect that the "deeper" elements, i.e. sections and sub-sections would be assessed as more specific than their super-elements. That is apparently not the case in this experiment.

4.3 Reliability of searchers' assessments

The data provides an opportunity to estimate the reliability of the searchers' assessments. In several cases, both whole articles and sections of articles have been assessed by a number of different searchers in relation to the same question. A study of these overlapping judgements shows a remarkable lack of consistency. Of the approximately 50 different articles which were judged by more than one searcher, there was full agreement between assessments in only five cases. Of these only one had more than two different assessments, and in three of these five cases the assessment was "J", i.e. "not relevant". In 10% of the cases both the categories "A" and "J" were included in the assessments of the same article, and more than 30% of the articles were judged to belong to five or more categories. On the section level, the inconsistencies were, if anything, even greater. Here too, the few times where agreement between searchers occurred, it was nearly always agreement on a "J" code.

A closer examination of one particular article the one which was assessed by the largest number of searchers, shows a greater than average agreement on the article level, with 7 assessments divided into 4 "A"s and 3 "B"s. On the section level, however, the four most frequently assessed sections, with 35, 21, 18 and 18 assessments, respectively, have their assessments spread over 6 to 9 of the 10 possible categories, and there is no significant difference in the degree of consistency between the "usefulness" and the "specificity" judgements. In the four sections with the highest number of assessments , assessment of specificity were distributed with 45% "very specific" , 41% "fairly specific" and 14% "marginally" or "not specific", whereas the same figures for usefulness were 54%, 33% and 13%, respectively.

There may be several explanations for this high degree of disagreement. 10 categories may be too many for the searchers to relate to in a consistent manner, or the four-part division of the two dimensions ("very", "fairly", "marginally" or "not") may be a difficult scale to interpret. Obviously a binary but possibly also a 5- or 7- part scale, used separately for the two dimensions, would have been easier to handle. The searcher's familiarity with the topics of the queries or their understanding of the material may of course also influence the reliability of the assessments. This may to an extent be clarified through an investigation of the searcher questionnaires, but since the disparity of assessments is universal over both articles and sections, such an investigation seems to be of dubious value. At any rate, judging of articles and sections seem to be an equally hard task, and the consistency problems calls for caution in the interpretation of the data presented above.

Discussion

One of the most alluring features of XML information retrieval has been the ability to perform segment-based indexing and document fragment retrieval (see e.g. [6]). The findings of our investigation support this contention; searchers appear to find more value in section-level than in article-level material, even if they still value the full article more highly in direct comparison. We have not studied the order in which assessments were made, this may throw more light on this apparent discrepancy.

A major limitation of this study is the lack of control with the data collection procedures. Data are collected from searchers around the world, with different backgrounds, pre-knowledge about topics, and information searching competence.

Although the simulated work task procedure was used, which aims at providing searchers with a common context, this method has important limitations. The most serious problem, which is also pointed out by [1], is that simulated work tasks should be adjusted somewhat towards the searchers' interest and backgrounds. This has not been the case in the INEX interactive track thus it is difficult, if not impossible, to say how background factors influenced the assessments.

Another weakness in the study is the definition of the XML elements sec, ss1 and ss2. We have not gone into the articles to see whether there are great discrepancies in the amount of text which constitute a section or a subsection, but have assumed them to be comparable with each other. In the main INEX experiments articles have been evaluated to a finer level of granularity, and the problems of evaluation have been even more apparent on the paragraph level; it would be interesting to see whether the same pattern of searcher assessment appears if they were exposed to this level, as a step towards defining the optimal level of text granularity for retrieval. So far, the research reported here only emphasize that this level is difficult to find.

Usefulness is a problematic concept to define; whereas specificity is a clearly defined term, having to do with the focus of topic treatment, usefulness is much vaguer. It may have been confused with or understood as specificity by the searchers. In the other INEX tracks the term "exhaustiveness" has been used to describe this dimension, and this term is easier to define. A better definition is particularly needed in experiments such as the one reported here, where the simulated work tasks were not tuned to fit the searchers performing the tasks, so that the concept of usefulness becomes a very theoretical notion.

A combined measure of relevance with so many alternatives as the one used in this experiment proves difficult for the searchers to relate to. In further experiments it might be fruitful to use another scale and resort to two separate assessments.

As shown in Section 4.3 there is a strong degree Inter-assessor agreement is always a problem, but rarely on the scale observed here.

References

1. Borlund, P.:. Evaluation of interactive information retrieval systems. Åbo: Åbo Akademi University Press (2000)

2. Finesilver, K, Reid, J.: User Behaviour in the Context of Structured Documents. In: Sebastiani, F. (ed.): Advances in Information Retrieval (ECIR 2003). Lecture Notes in Computer Science, Vol. 2633. Springer-Verlag, Berlin Heidelberg New York (2003) 104-119

3. Kekäläinen, J. & Järvelin, K.: Using graded relevance assessments in IR evaluation. Journal of the American Society for Information Science and Technology, 53(13) (2002) 1120-1129.

4. Kazai, G., Lalmas, M., Fuhr, N. Gövert, N.:. A report on the first year of the initiative for the evaluation of XML retrieval (INEX'02). Journal of the American Society for Information Science and Technology, 55(6) (2004) 551-556

5. Kazai, G, Lalmas, M, de Vries, A. P.: The Overlap Problem in Content-Oriented XML Retrieval Evaluation. In: Järvelin, K.,, Allan, J., Bruza, P., Sanderson, M. (eds.) Proceedings of the 27th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, Sheffield (2004) 72-79

6. Luk, R. W. P. et al.:. A Survey in indexing and searching XML documents. Journal of the American Society for Information Science and Technology, 53(6) (2002) 415-437

7. Malik, S., Tombros, A., Larsen, B.:. HyREX for INEX itrack. (2004) Available: http://inex.is.informatik.uni-duisburg.de:2004/tracks/int/internal/downloads/guide.pdf

8. Saracevic, T., Kantor. P., Chamis, A. Y., Trivison, D.: A Study of Information Seeking and Retrieving. I. Background and Methodology. Journal of the American Society for Information Science, 39 (3) (1988) 161-176

9. Tombros, A., Larsen, B. & Malik, S..: The Interactive track at INEX 2004. To be published in: Fuhr,, N., Lalmas, M., Malik, S., Szlavik, Z. (eds), INEX 2004 workshop pre-proceedings (2005)