The effect of task type on preferred element-types in an XML-based retrieval system

This is a postprint of an article published as Pharo, N & Krahn, A. (2011). The effect of task type on preferred element types in an XML-based retrieval system. Journal of the American Society for Information Science and Technology (September 2011), 62 (9), 1717-1726 (DOI: 10.1002/asi.21587)

Nils Pharo (corresponding author) a)

nils.pharo@jbi.hio.no

Astrid Krahn a)

Astrid.Krahn@skatteetaten.no

a) Oslo University College, PB St Olavs plass, 0130 Oslo, Norway

Abstract

This article examines the influence of task type on the users' preferred level of document elements (full articles, sections or subsections) during interaction with an XML-version of Wikipedia. We found that in general articles and subsections seemed to be the most valuable elements for our test subjects. For information gathering tasks this tendency was stronger whereas for factfinding tasks the sections seemed to play a more important role. We assume from this that users selected different information search strategies for the two task types. When dealing with factfinding tasks users seem in a higher degree to use one single element as an answer while they when they do information gathering they pick information from several elements.

1. Introduction

Although many information retrieval systems index and retrieve full documents they seldom index the entire document and its parts independently. Nevertheless in many situations only parts of documents will be relevant to a user's information need. In this article we report from a study of users' preferences with respect to document parts and how these differ for two types of tasks.

XML offers the possibility of indexing and retrieving semantically meaningful document parts (Luk, Leong, Dillon, Chan, Croft & Allan, 2002). An essential question is which types of elements in an XML-hierarchy are the most useful for users. In most XML-documents the markup is quite detailed – it is however not likely that users are interested in being presented with very small elements such as, e.g., titles or links. In the context of INEX (the Initiative for the Evaluation of XML retrieval) several authors have studied which element–types users prefer when searching in XML retrieval systems (Pharo & Nordlie, 2005; Kim & Son, 2006; Hammer-Aebi, Christensen, Lund & Larsen, 2006; Larsen, Tombros & Malik, 2006; Ramírez & de Vries, 2006; Pharo, 2008). In most of these studies a collection of scientific articles in the domain of informatics (IEEE) were used. A notable exception from the INEX-generated studies is the study by Balatsoukas and Demian (2010) on XML-coded documentation. It is reasonable to believe that genre will influence the way users read a document which in turn might influence which element-types are most appropriate for satisfying an information need. A second influencing factor might be the type of work task for which the information is needed (Byström &

Järvelin, 1995), whereas the users' topic knowledge (Marchionini, 1995) could be a third factor.

In this study we have investigated test subjects searching an XML-tagged version of Wikipedia. The data were collected in 2007 for the INEX Interactive Track. The search tasks analyzed in this study were categorized as two different types, information *gathering* and *factfinding*. As part of the experiment the test subjects were asked to fill in a questionnaire on, among other things, their knowledge about the topic, but a preliminary analysis of this data set made it impossible to draw any conclusions. The research questions examined in this study thus are:

1) What element types do people use when searching in an XML-version of Wikipedia?

2) How does the task type influence users element-type preferences?

This article is structured as follows: Section 2 presents related studies. In section 3 the experiment setup, the retrieval systems and the relevance scale used in this study are described. Section 4 reports our findings and section 5 contains discussion and conclusions.

1.1 Background

The work task is suggested to affect several factors during information retrieval processes, including the type and number of information sources used, the efforts invested by the user and the users' relevance assessments (Pharo, 2004). Work task can be defined in different ways, Byström and Hansen (2005), who have examined

the work task as a concept in library and information science, point out that work tasks can, on the one hand, be treated as objectively existing independent from the performer and with a clear defined outcome. On the other hand, the work task can be seen as something subjective which is defined by the performer. The authors also present different patterns of categorizing work tasks where one common criterion is the complexity of a task.

Byström and Järvelin (1995) categorized work task types according to complexity. They used five different categories ranging from *automatic information processing tasks* to *genuine decision tasks*. *Automatic information processing* is defined as being completely determinable while for *genuine decision tasks* neither the search-process nor the information requirements are known in advance. They find that with increasing complexity the need for *domain information* (known facts and theories within a problem's domain) and *problem solving information* (information on methods for handling problems) increase. During automatic information processing tasks, however, *problem information* (information directly solving the problem) is sufficient for solving the tasks, i.e. no additional contextual information are necessary.

Kim (2009) examines how different search strategies relate to different task types. She uses an information search strategy (ISS) scheme, based on the works of Belkin et al. (1993) and Cool and Belkin (2002), to identify information search strategies. The task types used are *factual task, interpretive task* and *exploratory task.* There seem to be marked differences in search behavior for *factual tasks* on one side and *interpretive* and *exploratory tasks* on the other side. Toms et al. (2007) points at several facets of tasks, including goals, domain, topic, process, structure and outcome,

listing common types of goals as “learning, fact-finding or information gathering” (p. 360). They examine the relationship between task type, using the categories above, and -structure (being parallel or hierarchical) and different aspects of query formulation and number of results viewed. Among other things they find significant differences in the number of queries and number of result pages viewed between different task types. Also task structure seems to influence query formulation in various ways.

Larsen, Malik and Tombros (2008) examined the degree of agreement between relevance-judgements in the INEX 2006/2007 Interactive Track and the distribution of relevance assessments for different task types at the INEX 2007 Ad Hoc track. They find indications that users found a larger proportion of relevant elements and a smaller proportion of non-relevant elements working on *information gathering* task whereas for *factfinding*-tasks the trend showed the opposite.

1.2. Users preferences of XML-elements

Pharo and Nordlie (2005) used data from INEX Interactive Track 2004, consisting of user interactions with a collection of XML-marked up computer science journals, and examined the effect of element type on relevance judgments. They found that the *section* was judged as the most relevant element both related to specificity and usefulness. However when a *section*-element and the *article*-element of the same document were assessed, the article-element was often assessed more relevant than the section-element.

Kim and Son (2006) compared user's interaction with two different XML-based retrieval systems, HYREX and Daffodil (both using the IEEE computer science journals collection). The main difference was that HYREX presented an unstructured result list, which means that different parts of one document could be found in different places in the result list, while Daffodil presented all retrieved parts of one document together. In both systems the *section*-element was seen and assessed more often than other element types, but in HYREX this tendency was stronger. The *section*-element also had highest score in relevance judgement in both systems.

Larsen et al. (2006) and Pharo (2008) analyzed data from INEX Interactive Track 2005. These studies show that most users first accessed the front matter-element (containing for a large part metadata) when examining a document. This element was also viewed most often. In contrast only a small part of the viewed fm-elements were relevance judged. The authors mention that the reason for this might be that users believe that they will be able to see the whole article by clicking on the title in the result list, but instead are lead to the fm-element. Most of the relevance-assessments were made for section-elements. The test subject in this study performed relevance-assessments of parts of documents more often than they did for the whole documents. Thus both studies conclude that elements are most useful for users, although whole articles score higher in relevance assessments once they are assessed.

Hammer-Aebi et al. (2006) examined user's interaction with XML-elements in a collection of guide books from *Lonely Planet*, which provide travel information on a wide range of destinations. They compare how whole articles are relevance-assessed compared to elements. They found that the major part of the exact-assessments was

made on elements, especially elements on the coarser levels of granularity, i.e. element types which might contain several other elements. Interestingly, the authors found that users did not care too much about seeing the information in context of the whole document.

Ramírez and de Vries (2006) examined if topic knowledge and task type influence which element type users prefer. They found that users working with simple tasks preferred elements on a high level of granularity. Working with narrow tasks users considered *sections* more often relevant compared to when they were working with broad tasks. Interestingly, in the latter group both articles and subsections were more often judged *relevant* when compared to narrow work tasks. Users with high topic knowledge considered elements on a fine level of granularity more useful than users with less topic knowledge.

Kamps and Larsen (2006) asked the topic creators for the INEX Ad Hoc Track what kind of information they expected to be useful for matching their topics. Users with high topic familiarity often requested specific information which they expected to be short. They did not consider finding all relevant information as necessary. Users who were looking for more comprehensive information wished to be presented with all relevant information units and expected that it would be interesting to read several documents.

2. Method

We used data collected for the INEX Interactive Track 2006/2007 1). In this section we describe the experimental setup.

The test subjects were recruited from research groups at 8 different universities, in total 88 participants were recruited, of whom data from 69 were used in our study. A large majority of the participants were students. In the experiments two different systems were compared, one passage retrieval system and one element retrieval system. Two sessions in each of the systems were performed (i.e. four tasks per participant). The test subjects could choose between three different tasks for each of the four sessions. In our analysis we have only investigated data from the element retrieval system. Each session lasted a maximum of 15 minutes. Before the experiment participants filled out a questionnaire giving background data, such as age, gender, spoken languages and field of study. Before each of the tasks they filled out questionnaires on their topic knowledge and after each task and in the post-experiment questionnaire they answered system-related questions. The choice/relevance assessments of elements were collected from the transaction logs. Below we present the main components of the experiment and emphasise the variables used in our data analysis.

2.1 The search system

The participants were asked to search for information for a number of search tasks in an XML-version of Wikipedia using the experimental IR system Daffodil (Figure 1) (Malik, Tombros and Larsen, 2007). The collection contains almost 660 000 Wikipedia articles (Denoyer & Gallinari, 2007).

Insert Figure 1 System search interface

Figure 1 shows the search interface and a result list from Daffodil. The system has a simple search box where one or more search terms are entered. In the result list all relevant hits from one document are clustered with the titles of the relevant sections / subsections appearing below the title of the whole document. The user can choose to access the document via the document title which leads him to the start of the document. Alternatively he can enter the document via one of the listed (sub)sections, which will lead him directly to this paragraph with the option to scroll up and down in the whole document.

Figure 2 shows the document view of Daffodil. On the left hand side is the table of contents of the document, where users can enter other potentially interesting document parts. Four document hierarchy levels, represented by XML elements, are available for individual inspection, these are (from the coarsest level of granularity to the finest): the full *article*, *sections*, *subsections*, and *subsubsections* (all elements were tagged "section" in the collection – see Figure 3). The size of the elements differs between documents.

Insert Figure 2 System document view

Insert Figure 3 Generic structure of Wikipedia XML collection

Our first research question seeks to reveal what element types are most useful for solving the tasks. We use three different measures: the number of elements of different types inspected, the share of the inspected elements on the different levels of

granularity which are relevance assessed, and the values of the participants' relevance assessments of the different element types.

The combination of these measures is chosen to get an impression of both what users think (by analyzing the values of relevance assessments) and what users actually do (from the numbers of elements assessed and elements inspected).

2.2 Tasks

The search tasks were given as simulated work task situations (Borlund, 2003) and are presented in Appendix 1. The participants were asked to spend a maximum of 15 minutes per task. In all, there were three different task types, based on the work of Toms et al. (2006); *decision making*; *information gathering* and *factfinding*. In this study we only examined the two latter ones. Factfinding was defined as tasks "…where the objective is to find 'specific accurate or correct information or physical things that can be grouped into classes or categories for easy reference'" (Malik et al., 2007, p. 395). Information gathering was defined as tasks "…where the objective is to collect miscellaneous information about a topic" (Malik et al. 2007, p. 395).

The reason for omitting decision making tasks from our analysis was that these tasks were quite mixed with regard to complexity. Most of the information gathering tasks were quite indeterminable while most of the factfinding tasks were very structured and well-defined. In order to have clear contrasts we chose to concentrate the analysis on these two categories. Since the original task 7 was more complex than the other factfinding tasks and the original task 10 less complex than the remaining information gathering tasks these two tasks were also excluded from the analysis. As we mentioned earlier no sessions performed in the passage retrieval system were used.

Research question 2 seeks to analyze the interdependency between task type and the inspection of and relevance assessments of the different element types. Thus we have a twofactorial design with two independent variables, element types and task types.

2.3. The relevance scale

One of our main measures is relevance assessment; the scale used in the experiments was based on the work by Pechevski (2006). The intention of which is to use a scale that also takes into account the hierarchical structure of XML documents, in our case that articles contain sections, sections contain subsections and so on. The relevance scale used measured two aspects of relevance, topical relevance and specificity. Topical relevance is here measured on a three-degree scale, inspired by the experiments in the IR community to distinguish between highly relevant and relevant documents (Järvelin & Kekäläinen, 2000; Voorhees, 2001), our scale taking the values relevant, partially relevant and non-relevant. The specificity dimension was meant to indicate how much context was needed to understand the information in an element, e.g., the test subject should indicate whether the section as a part independent of its mother (article) contained an appropriate amount of information. Ideally the retrieved element should be self-contained, but sometimes the element is "too broad", meaning that it also contains information not related to the query. In other cases the element itself is "too narrow", i.e. its content is relevant to the query, but additional information from surrounding elements is needed. (Malik et al. 2007). Thus in the user guidelines five possible relevance scores were presented to the participants:

- **Relevant, but too broad**, contains relevant information, but also a substantial amount of other information

- **Relevant**, contains highly relevant information, and is just right in size to be understandable

- **Relevant, but too narrow**, contains relevant information, but needs more context to be understood

- **Partial answer**, has enough context to be understandable, but contains only partially relevant information

- **Not relevant**, does not contain any information that is useful in solving the task

The participants were asked to relevance assess every element they read, but there was no system mechanism included to force them to add their assessments.

2.4. Logs

The Daffodil system provides rich transaction log data and in our analysis we used log data on the elements used and the relevance assessments per element. To help in our analysis the hierarchical structure and titles of individual elements were recorded in the logs.

2.5 Analysis

We use frequency distribution of our observations and entered the data into cross-tables. To measure the significance of our cross-tabular analysis we conducted Chi-square tests.

3. Results

Data from 69 participants performing 87 tasks were analyzed. 29 of the task sessions were related to the task type *information gathering* and 58 sessions were related to *factfinding*, Table 1 presents the distributions of participants per task. In all, the participants looked at 1060 elements on different levels of granularity and they assessed the relevance of 729 elements.

Insert Table 1

3.1. Relevance-judgments independent of task type

The first section reports how our test subjects deal with the different element types in general, independent of task type. We used numbers of element views, the percentage of relevance assessments per element type viewed and the value of the relevance assessments as our measures of element type use.

Insert Table 2

Table 2 shows the distribution of viewed element types and how many of those are relevance assessed. We see that the number of element views is highest for section elements, closely followed by articles whereas considerably fewer subsections and subsubsections are inspected by the participants. Since the distribution of the elements in the collection is not known 2), however, these numbers has to be interpreted with caution. For example, not every article has subsections, which might explain the lower views on this element type compared to sections. On the other hand it is obvious that there are more sections than articles in the collection, because most

articles consist of several sections. Nevertheless, the fact that article-elements are viewed almost as often as section-elements indicates that full Wikipedia articles are quite important for solving users' information needs. Below we shall look at how this differs for the two types of tasks.

The share of relevance assessed elements compared to the viewed elements is quite similar for sections (65.1 %) and subsections (68.6 %). This indicates that the differences with respect to element views are not related to user preferences but rather are a result of their distribution in the collection. The element types for which the proportion of relevance assessed elements differ the most from viewed elements are article and subsubsection. Articles are assessed most often whereas subsubsections are assessed seldom.

Insert Table 3

We wanted to break down the relevance assessments to look at the topical relevance dimension and the specificity dimension separately. In Table 3 the *relevant*-column includes all elements which are judged *relevant*, *too broad* and *too narrow*. We also excluded subsubsections because only 9 relevance assessments were made on this element-type (3 not relevant, 4 fully relevant, 2 too narrow).

As we see the relevance assesment differs between the different element types ($p<0.001$), subsection has the highest proportion of topical *relevant*-assessments, while article has the lowest. It is somewhat surprising that articles are more often

judged *not relevant* compared to the other element types. A possible reason is that users, when not finding anything relevant in an article, do not see the need to relevance-assess each sub-element, but rather assess the document as a whole as *not relevant*. When parts of a document are relevant it makes more sense to relevance-judge the single document parts separately. That might also be the explanation for why articles are being relevance-assessed more often than other elements.

Insert Table 4

Table 4 shows relevance-assesments related to *specificity*, thus we have excluded the partially relevant and not-relevant assessments from the analysis. Also here we see significant differences between the element types ($p<0.01$). The subsection has the highest proportion of fully-*relevant*–assessments, whereas the section-element has the lowest proportion. This is consistent with the findings of Kamps and Koolen (2007) who found that subsections and articles more often fitted with a relevant passage than sections. Possibly the section element often contains too much information when the user needs a short, concise answer but is too small to satisfy users who look for more comprehensive information.

The results from relevance-assessments suggest that subsections are of most value to users. Subsections are considered as most relevant both related to specificity and to topical relevance. The subsections are, however, inspected quite seldom and one could argue that they are only viewed when the user expects them to be especially interesting. On the other hand, we know that not all articles have subsections, and that the proportion of relevance-assessed subsections is slightly higher than for sections.

This makes it natural to assume that the low percentage of inspected elements for this element type is a result of the distribution of elements in the collection.

We have also seen that sections are viewed relatively more seldom than article-elements and that compared to full articles and subsections the section scores lowest in relation to specificity and holds the middle position for topical relevance.

In conclusion these results suggest that subsections and articles are the most valuable elements for users, but that also sections seem to be important. Subsubsections on the other hand, are accessed seldom and have a very low percentage of relevance-assessments. That indicates that users do not consider them as independent information units.

3.2 The effect of task type

This section presents which element types the participants accessed and assessed when dealing with different task types. As mentioned above we had exactly twice as many factfinding sessions as information gathering session due to the distribution of tasks among test subjects and the two different IR systems. Information gathering sessions generated 14.5 assessements per session, whereas there were only 11 assessments per factfinding-session. This is an indication that our sample of information gathering sessions resulted in more user activity than the factfinding sessions.

Our hypothesis was that users dealing with information gathering–tasks prefer larger information units compared to those dealing with factfinding-tasks. This was based on

the assumption that facts could be derived from smaller parts of the document. We thus predicted that information gathering-tasks should result in a higher number of viewed elements and a larger proportion of relevance-assessments for larger element types. Moreover more elements on a coarse level of granularity (i.e. larger parts of the document) should be assessed more often as relevant for information-gathering-tasks than for factfinding-tasks. For the same reason we expected more too narrow assessments and fewer too broad assessments for information-gathering tasks.

Insert Table 5

Insert Table 6

Table 5 shows the distribution of elements inspected for factfinding and information gathering and reveals significant differences for the two task types ($p<0.001$). Test subjects dealing with *information-gathering* tasks, in contrast to our expectations, inspected a much larger share of small element-types than for the *factfinding* tasks. However we see that the smallest element type subsubsection was relevance-assessed seldom compared to the other elements (Table 6).

It is remarkable that, for *information-gathering* tasks, participants relevance-assessed a lower proportion of the viewed elements than for *factfinding* tasks. With respect to article-elements, however, the proportion of relevance-assessed elements is the same for *factfinding* and *information gathering* tasks. This might indicate that users, even if they look at many elements on a fine level of granularity (i.e. small elements) when dealing with *information gathering*, consider the article as their context of reference.

Thus they might not bother relevance-assessing every minor element they read. The difference in assessments between the two task types is remarkable, when we compare sections and subsections we see that during information gathering tasks our participants have assessed a much higher share of subsection elements (66.7 % compared to 55.9 % of the sections). Above we made note of the seemingly importance of subsections to users, and now we have found that this element is particularly useful for information gathering tasks.

Insert Table 7

Table 7 shows the distribution of relevance assessments for topical relevance for our two task types. The Chi-square test reveals significant differences between the element types for the factfinding task ($p<0.001$), but not for the information gathering tasks. This means that the task type clearly influences the relevance assessments, and that for factfinding tasks element granularity influences the assessment. Comparing the two task types we see that the test subjects in general consider a much higher share of the elements as relevant for *information gathering*-topics than for *factfinding*-tasks. That seems natural taking into account that tasks of the *factfinding* type are quite narrow by definition. The participants are asked to find very specific information – thus if the required information is not found the element will be judged as *not relevant*.

During *factfinding* the section element has the highest percentage of relevant-judgements, which indicates that this is the most useful document part. Also these results contradict our hypothesis, but are in line with our reasoning above. This is consistent with the findings from Ramírez and de Vries (2006) who found that users

dealing with narrow topics (in the INEX IEEE journal collection) had a stronger tendency to prefer sections than users dealing with broad topics.

A possible reason why users seem to be more indifferent with respect to element type for information gathering is that they use the whole document as their context of reference while jumping to different paragraphs they consider as possibly interesting.

Insert Table 8

Table 8 shows relevance judgements related to specificity for the different task types. Also here we only have significant differences ($p<0.05$) between element types for factfinding, but not for information gathering tasks.

In total, *factfinding* task sessions have resulted in more too broad assessments, but surprisingly also in more too narrow assessments than those initiated by information gathering tasks. In particular, for the section-element there are more too narrow-evaluations for *factfinding* tasks than the *information gathering* tasks. A reason can be that users performing *factfinding* use the section element mainly as a single answer whereas when they perform *information gathering* tasks they use them as part of an answer. This explains why our test subjects more often wish to obtain additional context in the category *factfinding*. These findings are supported by Kim's (2009) study of general web search behaviour. In factual tasks, which are similar to our category *factfinding,* search strategies using the mode *specify* are dominating. *Specify* is defined as "*Search for an item"*. That means users are looking for one special item as opposed to *recognize* strategies where users are *looking around in an item.*

Strategies using this latter mode are dominating in *interpretive* and *exploratory* tasks which are comparable to our *information gathering*-category.

Concluding we see that participants dealing with *information gathering* have a tendency to pay more attention to the very small elements compared to when they do *factfinding*, but still they seem to consider articles as important, which is shown by the comparably high proportion of relevance-assessments of articles. Probably the article is considered as the main information unit and thus is browsed for partial answers. Wikipedia-articles often deal with several different aspects of the same topic, which might make it natural to pick information from several document parts. On the other hand we see that participants doing *factfinding* seem to have a tendency to use larger element types compared to their preferences when performing *information gathering*. During factfinding relatively more sections and articles are viewed than in *information gathering* (cf Table 5). Also, a relatively larger proportion of sections are relevance-assessed during *factfinding*. Moreover sections have the highest proportion of topical *relevant* elements in *factfinding* and the, by far, lowest percentage of *not relevant* assessments. This suggests that section is quite an important element for *factfinding* topics even if it often is considered *too broad* or *too narrow*. As mentioned it is likely that users in *factfinding* consider the elements they look at more as a whole answer, whereas working with information gathering means more 'picking of information bits here and there'. This is supported by the fact that the *factfinding* participants relevance-assess a much higher proportion of the elements they look at.

4. Discussion and conclusions

Our article presents the analysis of users' preference of elements in Wikipedia articles. Although the presentation format and search interface from Daffodil is different from the ordinary Wikipedia interface, we believe our findings can be of great value for the structuring of encyclopaedic texts for information retrieval. The design of the experiment must, however, be taken into consideration when discussing the implications of our findings on system design.

The relevance scale in INEX 2006/2007 consisted, as mentioned, of two dimensions, thus we chose to separate the results according to these dimensions, topical relevance and specificity. Of course this separation is somewhat artificial. Spink, Greisdorf and Bateman (1998) found that users associate both *too broad* and *too narrow* with the term *partially relevant*. It is likely that the participants in our study not strictly distinguished between these categories.

In our analysis we have used the relevance assessments to signify what elements users prefer. When searching to solve real (as opposed to simulated) tasks, however, users will often *use* only a small share of the relevant documents or document parts. Thus we need to perform studies of real users that perform real tasks in order to learn more about the optimal solutions for XML retrieval. In the 2010 INEX iTrack experiment, which is yet to be analysed, users interacted with a collection of book surrogates, simulating a digital book store. In these experiments, which also included relevance assessments, the users were asked to add to a shopping cart the books they would have bought. This is one possible solution to get more realistic data on user preferences in simulated experiments.

Which elements users choose to look at will always be influenced by the user interface of the retrieval system. In Daffodil the article element might have been favoured in regard to the number of elements viewed and assessed. In the result list users find the article title on top of all results from one document. As mentioned Kim and Son (2006) found that the preference for sections was stronger in HYREX where different parts of a document could be presented in different places in the result list. In 2005, when Daffodil was used, Pharo (2008) and Larsen et al. (2006) found that the fm–element which was in the same place as the document title in 2006 - was the most viewed element. Thus it seems reasonable to assume that the Daffodil interface promotes the choice of full articles as entry points, this is most probably independent of document genre (i.e. journal article or encyclopaedic text). Moreover, when test subjects wanted to go back to the result list after examining a document they were forced to close it actively, this probably also serves as a reminder to relevance assess the article. Contrary to this the user's "leaving" of a section/sub-section did not, in the same way, force him/her to perform a relevance assessment.

The hierarchical structure of XML documents also needs to be taken into account when considering the results of our analysis. Since articles, sections and subsections are overlapping elements, we cannot be certain that when, e.g., a user assesses a section that it is not in fact one of its subsections he/she finds relevant. We therefore have to trust that the user indeed has followed the instructions to judge the relevance of all elements they read.In our study we also have assumed that elements on a coarser level of granularity are larger than elements on a finer level of granularity. Of course there are large variations in the size of elements on the different levels of

granularity. However it seems reasonable to assume that elements on a coarse level of granularity in general are more comprehensive and more self contained than elements on a finer level of granularity.

Our results show that users in this collection considered subsections and articles as the most valuable elements, in contrast to earlier studies, many of which conclude with sections being the most interesting element (Larsen et al., 2006; Pharo, 2008; Kim & Son, 2006). This is an indication that users' preferences differ across different document genre, since the previous experiments used a similar interface (Daffodil) on computer science articles. However the preference of element-types seems to a large degree to depend on task type. For information gathering tasks users seem to be "open" for all element types while for factfinding they seem to regard section type as especially useful. We believe that this is a result of different information search strategies. When performing *factfinding* tasks the user skims the documents for one information unit containing a very specific piece of information, i.e. *problem information* (Byström & Järvelin, 1995) which directly helps the user in solving the task. Probably for *information gathering* tasks users in higher degree need *domain information* as defined by Byström and Järvelin (1995) in contrast to factfinding task where problem information is most important. Thus when the user picks various information on different aspects of the topic, he still uses the whole document as his main information unit since this is what provides him with the necessary context.

For the development of new retrieval systems our findings suggest that for *factfinding* tasks it is reasonable only to present the most relevant results while for *information gathering* it is desirable to get as much information as possible.

Footnotes

1) The 2006 iTrack data collection was delayed and did not take place until spring 2007

2) The corpus' XML structure is very intricate. According to Denoyer and Gallinari (2007), the 659388 documents contain approximately 52 million elements

Acknowledgements

INEX was partly funded by DELOS, a network of excellence in digital libraries. We would like to thank the participants of the 2006 interactive track. We would also like to thank our anonymous reviewers for good advice.

Reference List

Balatsoukas, P., & Demian, P. (2010). Effects of granularity of search results on the relevance judgement behaviour of engineers: building systems for retrieval and understanding of context. Journal of the American Society for Information Science and Technology, 61(3), 453-467.

Belkin, N. J., Marchetti, P. G., & Cool, C. (1993). BRAQUE: Design of an interface to support user interaction in information retrieval. Information Processing and Management, 29(3), 325-344.

Borlund, P. (2003). The IIR evaluation model: a framework for evaluation of interactive information retrieval systems. Information Research, 8(3). Retrieved November 17, 2010, from http://informationr.net/ir/8-3/paper152.html.

Byström, K. & Hansen, P. (2005). Conceptual framework for tasks in information studies. Journal of the American Society for Information Science, 56(10), 1050-1061.

Byström, K. & Järvelin, K. (1995). Task complexity affects information seeking and use. Information Processing and Management, 31(2), 191-213.

Cool, C. & Belkin, N. J. (2002). A classification of interactions with information. In Emerging Frameworks and Methods: CoLIS 4 (pp. 1-15). Greenwood Village, Col, Libraries Unlimited.

Denoyer, L. & Gallinari, P. (2007). The Wikipedia XML corpus. Lecture notes in computer science, 4518, 12-19.

Hammer-Aebi, B., Christensen, K. W., Lund, H., & Larsen, B. (2006). Users, structured documents and overlap: interactive searching of elements and the influence of context on search behaviour. In In IIiX: Proceedings of the 1st international conference on Information interaction in context (2006), (pp. 46-55). ACM Press New York, NY, USA.

Järvelin, K., & Kekäläinen, J. (2000). IR evaluation methods for retrieving highly relevant documents. In Belkin, N., Ingwersen, P., & Leong, M.-K. (Eds.), Proceedings of the 23th Annual International ACM SIGIR conference on research and development in information retrieval (pp. 41–48). New York: ACM Press.

Kamps, J. & Koolen, M. (2007). On the relation between relevant passages and XML document structure. In A. Trotman, G. Shlomo & J. Kamps (Eds.), Proceedings of the SIGIR 2007 Workshop on Focused Retrieval, Seattle, (pp. 28-32). Dunedin, University of Otago.

Kamps, J. & Larsen, B. (2006). Understanding Differences between Search Requests in XML Element Retrieval. In A. Trotman & G. Shlomo (Eds.), Proceedings of the SIGIR 2006 Workshop on XML Element Retrieval Methodology (pp. 13-19). Dunedin, University of Otago.

Kim, H. & Son, H. (2006). Users Interaction with the Hierarchically Structured Presentation in XML Document Retrieval. Lecture notes in computer science, 3977, 422-431.

Kim, J. (2009). Describing and predicting information-seeking behavior on the Web. Journal of the American Society for Information Science and Technology, 60(4), 679-693.

Larsen, B., Malik, S., & Tombros, A. (2008). A Comparison of Interactive and Ad-Hoc Relevance Assessments. Lecture notes in computer science, 4862, 348-358.

Larsen, B., Tombros, A., & Malik, S. (2006). Is XML retrieval meaningful to users?: searcher preferences for full documents vs. elements. In E. Efthimiadis, D. Hawking, & K. Järvelin (Eds.), (pp. 663-664). ACM New York, NY, USA.

Luk, R. W. P., Leong, H. V., Dillon, T. S., Chan, A. T. S., Croft, W. B., & Allan, J. (2002). A survey in indexing and searching XML documents. Journal of the American Society for Information Science and Technology, 53(6), 415-437.

Malik, S., Tombros, A., & Larsen, B. (2007). The Interactive Track at INEX 2006. *Lecture notes in computer science*, *4518*, 387-399.

Pehcevski, J. (2006). Relevance in XML retrieval: The user perspective. In Proceedings of the 29th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (pp. 35–42). NewYork: ACM Press.

Pharo, N. (2004). A new model of information behaviour based on the search situation transition schema. Information Research, 10(1). Retrieved November 17, 2010, from http://informationr.net/ir/10-1/paper203.html.

Pharo, N. (2008). The effect of granularity and order in XML element retrieval. Information Processing & Management, 44(5), 1732-1740.

Pharo, N. & Nordlie, R. (2005). Context matters: An analysis of assessments of XML documents. Lecture notes in computer science, 3507, 238-248.

Ramírez, G. & de Vries, A. (2006). Relevant contextual features in XML retrieval. In Ruthven, I. et al. (Eds.), Information interaction in context: International symposium on information interaction in context IiiX 2006, Copenhagen, (pp. 56-65). Copenhagen: Royal School of Library and Information Science.

Spink, A., Greisdorf, H., & Bateman, J. (1998). From highly relevant to not relevant: examining different regions of relevance. Information Processing and Management, 34(5), 599-621.

Toms, E. G., O'Brien, H., Mackenzie, T., Jordan, C., Freund, L., Toze, S. et al. (2007). Task effects on interactive search: The query factor. Lecture notes in computer science, 4862, 359-372.

Vorhees, E.M. (2001). Evaluation by highly relevant documents. In D.H. Kraft et al. (Eds.), Proceedings of the 24th Annual International Conference on Research and Development in Information Retrieval of the Special Interest Group on Information Retrieval, Association for Computing Machinery (SIGIR 01) (pp. 74–82). New York: ACM.

Information gathering tasks		Factfinding tasks	
Task#	Participants	Task#	Participants
Task 5	19		
Task 6	15		
Task 8	24		
		Task 9	11
		Task 11	11
		Task 12	7
Total	**58**	**Total**	**29**

Table 1 Distribution of participants per task

	Element-views	Proportion of relevance-assessed elements
Article	**39.9%** (423)	**77.3 %** (327)
Section	**43.3%** (459)	**65.1%** (299)
Subsection	**12.9%** (137)	**68.6 %** (94)
Subsubsection	**3.9%** (41)	**22.0%** **(**9)
Total	**100%** (1060)	**68.8%** (729)

Table 2 Distribution of viewed element types and relevance assessments

	Relevant	Partially relevant	Not relevant	Total
Article	**53,2%** (174)	**13,5%** (44)	**33,3%** (109)	**100%** (327)
Section	**58,9%** (176)	**21,1%** (63)	**20,1%** (60)	**100%** (299)
Subsection	**64,9%** (61)	**17,0%** (16)	**18,1%** (17)	**100%** (94)
Average	**57,1%** (411)	**17,1%** (123)	**25,8%** (186)	**100%** (720)

Table 3 Distribution of topical relevance on elements

	Fully relevant	Too broad	Too narrow	Total
Article	**69 %** (120)	**25.9 %** (45)	**5.2 %** (9)	**100 %** (174)
Section	**64.2 %** (113)	**22.2 %** (39)	**13.6 %** (24)	**100 %** (176)
Subsection	**75.4 %** (46)	**8.2 %** (5)	**16.4 %** (10)	**100 %** (61)
Average	**67.9 %** (279)	**21.7 %** (89)	**10.5 %** (43)	**100 %** (411)

Table 4 Distribution of relevance, according to level of specificity

	Factfinding	Information gathering
Article	**44.1%** (282)	**33.6%** (141)
Section	**48.0%** (307)	**36.2%** (152)
Subsection	**7.8%** (50)	**20.7%** (87)
Subsubsection	**.2%** (1)	**9.5%** (40)
Total	**100.0%** (640)	**100.0%** (420)

Table 5 Viewed elements per task type

	Factfinding	Information gathering
Article	**73.6%** (220)	**75.9%** (107)
Section	**69.7%** (214)	**55.9%** (85)
Subsection	**72 %** (36)	**66.7 %** (58)
Subsubsection	**100 %** (1)	**20 %** (8)
Average	**73.6%** (471)	**53.8%** (258)

Table 6 Assessed elements per task type

	Factfinding				Information gathering			
	Relevant	Partially relevant	Not relevant	Total	Relevant	Partially relevant	Not relevant	Total
Article	**49.1%** (108)	**13.2%** (29)	**37.7%** (83)	**100%** (220)	**61.7%** (66)	**14.0%** (15)	**24.3%** (26)	**100%** (107)
Section	**57%** (122)	**23.4%** (50)	**19.6%** (42)	**100%** (214)	**63.5%** (54)	**15.3%** (13)	**21.2%** (18)	**100%** (85)
Subsection	**47.2%** (17)	**22.2%** (8)	**30.6%** (11)	**100%** (36)	**75.9%** (44)	**13.8%** (8)	**10.3%** (6)	**100%** (58)
Average	**52.6%** (247)	**18.5%** (87)	**28.9%** (136)	**100%** (470)	**65.6%** (164)	**14.4%** (36)	**20.0%** (50)	**100%** (250)

Table 7 Distribution of topical relevance on task type

	Factfinding				Information gathering			
	Fully relevant	Too broad	Too narrow	Total	Fully relevant	Too broad	Too narrow	Total
Article	**66,7 %** (72)	**28,7 %** (31)	**4,6 %** (5)	**100 %** (108)	**72,7 %** (48)	**21,2 %** (14)	**6,1 %** (4)	**100 %** (66)
Section	**61,5%** (75)	**22,1%** (27)	**16,4%** (20)	**100%** (122)	**70,4%** (38)	**22,2%** (12)	**7,4%** (4)	**100%** (54)
Subsection	**70,6%** (12)	**11,8%** (2)	**17,6%** (3)	**100%** (17)	**77,3%** (34)	**6,8%** (3)	**15,9%** (7)	**100%** (44)
Total	**64,4%** (159)	**24,3 %** (60)	**11,3 %** (28)	**100 %** (247)	**73,2 %** (120)	**17,7 %** (29)	**9,1 %** (15)	**100 %** (164)

Table 8 Distribution of relevance specificity on task type