XML and Context
Structural Features Relevant to Search Tasks

Georgina Ramírez and Arjen P. de Vries
CWI, PO Box 94079, 1090GB Amsterdam, The Netherlands
{georgina,arjen}@cwi.nl

1. INTRODUCTION

The work presented in this paper is part of a research project that studies and develops an experimental approach to correlate user needs and IR strategies on the use of structural information. Its goal is to investigate the use of structural features for effective information retrieval and to define a model to link search tasks to specific retrieval strategies.

In this paper, we define a first classification of search tasks in structured document collections based on two dimensions: the type of information a user is seeking for (informational or resource) and the user familiarity with the structure of the documents (high, some or none).

We then classify the INEX[1] 2004 topics into the defined task categories, and analyse the relevance of several structural features regarding each of these tasks for the INEX collection. We investigate the following research questions:

1. Can we identify a (measurable) dependency between a topic's task type and the structural aspects of the topic's relevant documents?

2. Which structural features, if any, are relevant to the different degrees of knowledge a user might have on the structure of the documents?

2. USER TASKS IN STRUCTURED IR

2.1 Motivation

Previous studies have categorised search tasks or query types in different areas of information retrieval, such as question answering (e.g. [7]), web search (e.g. [3], [6]), or information systems in general (e.g. [1], [2]).

These studies use different contextual factors to classify information needs. For instance, user studies in digital libraries classify user tasks according to the amount of information needed by the user (e.g., specific or collecting information), the aim of the information seek (e.g., simple fact questions, decision questions, comparison questions), or the knowledge the user has on the topic, to mention some.

We believe that most of these classifications could be directly applied or adapted to structured information retrieval. However, this new scenario may require extension of the classifications to include the degree of knowledge a user might have on the structure of the documents. We argue that the information that a specialist like a librarian has about the structural components of a collection most likely differs from that of an unexperienced end user. This type of contextual information may also be important for an IR system to be able to distinguish different types of information needs and treat them accordingly.

As an example, consider a user interested in finding a book review that discusses *context in IR*. A user familiar with the structure of the INEX collection (e.g. the librarian) could know that generally book reviews appear in sections of documents titled "new books", "book review" or "bookshelf". He or she might then pose the following query[2]:

//article[about(.//atl, "new book" "book review"
bookshelf)]//sec[about(., context IR)] [3]

While a user less familiar with the INEX collection would probably simply ask:

//sec[about(., book review context IR)]

The structural constraints of the librarian may help the retrieval system to perform a better search and maybe even reduce the search space and therefore should be treated in a *stricter* way. However, in the case of the unexperienced end user, we do not want the retrieval system to use the structural constraints at all. If we restrict the search and use only the sections to find the query terms we might not find the desired information, since the exact phrase "book review" might appear neither in the section title nor in the section body.

We believe that the knowledge users have about the structure of the documents should be an important contextual factor when choosing a retrieval strategy. The following section investigates this hypothesis using the INEX collection, by analysing the relationship between the relevant documents and their structural features.

2.2 A basic taxonomy

We first define a rather straightforward classification of user tasks that involve searching structured document collections. The search tasks are classified using two different dimensions: *task type* and *collection familiarity*. The first one, similarly used in other categorizations (e.g. [7]), classifies tasks according to the type of information searched for: in our case, an *informational* task entails collecting information about a topic (even if this information is very specialized in content) and a *resource* task entails looking for a specific type of resource about a topic (e.g., reference, book review, algorithm, . . .).

[1] INEX is briefly described in Section 2.3.

[2] Queries are expressed in NEXI [9], INEX's query language. NEXI is a subset of XPath extended with an *about* clause used for ranking purposes.

[3] atl⇒article title, sec⇒section.

The second dimension, *collection familiarity* classifies tasks according to the user's degree of knowledge of the collection's structure. We define three different levels: *high*, when the user has detailed knowledge of the collection's schema (even if this is not complete), *some*, when the user knows some tags or has a general idea of what can be found, and, *none*, when the user is not aware of the structure and therefore uses plain text queries.

Table 1 shows the resultant user search task taxonomy.

Table 1: User search task taxonomy. The figures inidicate the number of INEX topics belonging to each category.

	Collection familiarity		
Task type	**None**	**Some**	**High**
Informational	A (17)	B (6)	C (3)
Resource	D (8)	E (5)	F (8)

2.3 INEX topics classification

The *Initiative for the Evaluation of XML retrieval* (INEX) [4] is a benchmark for the evaluation of XML retrieval. The collection provided to the participants is a subset of IEEE Computer Society publications. The participants are responsible for creating a set of topics and for assessing the relevant components for each of these topics. The relevance judgement is given by two different dimensions: exhaustivity (E) and specificity (S). A four-level scale (0-3) is used in both dimensions to specify the degree of relevance for each of them. More information about this process can be found in [5]. In this paper, to keep it concise, we use only the information of elements assessed as $(E, S) = (3, 3)$, the ones considered *highly relevant.* INEX defines two main tasks: content-oriented (CO) and content-and structure (CAS). The queries for the CO task are free text queries for which the retrieval system should retrieve relevant XML elements of varying granularity, while the queries for the CAS task contain explicit structural constraints. We consider both CO and CAS queries as expressions of information needs and treat them identically, i.e., we apply the same rules to classify them into our taxonomy.

We manually classified all INEX 2004 topics into the six categories defined in Subsection 2.2. We did so by looking at the *narrative* field of the topic description. The narrative field is a natural language description of the information need where the user describes the information need, the context and motivation for the search, and what makes an element relevant or not. The narrative field is the one used later to define the relevance of the elements.

Topics were divided into *informational* and *resource* according to the main goal of the narrative description. Topics containing sentences like "I am looking for information about..." or "I am interested in articles about..." were clustered into the *informational* class, while topics containing specific constraints on the type of information searched, such as "Find experimental results on...", "I am looking for definitions of..." were classified as *resource* tasks.

Since the user's familiarity with the collection is not mentioned in the INEX topic description, we established the following rule to classify topics according to this familiarity: Topics which do not contain any structural restriction (i.e., CO topics) were classified as *none* (user does not have knowledge of the structure).

We defined a common (and intuitive) tag set consisting of article, section (sec), paragraph(p), abstract(abs) and body(bdy). Topics with structural restrictions containing these common tags were classified as user tasks where the user has *some* knowledge of the collection schema, based on the argument that any collection of scientific articles would be expected to contain these type of elements. In other words, using these elements does not require detailed knowledge of the structure, i.e., it is not collection specific.

Topics that refered to tags outside the common tag set were classified as involving a user with *high* knowledge of the collection structure because the use of specialized elements (e.g., in the INEX case, atl, st, bb, fm, fig)[4] indicates that the user is very familiar with the specific structure of the INEX collection.

From all INEX 2004 topics (35 CAS and 40 CO) 14 were not assessed and 13 had no elements assessed as highly relevant. The remaining 47 topics were classified into our six categories. The number of topics belonging to each category are shown in Table 1.

3. STRUCTURAL FEATURES RELEVANT TO USER TASKS

This section analyses three different features from the INEX 2004 relevance assessments and determines if these exhibit differences related to (1) the task type categories (informational-resource), (2) the collection familiarity categories, and (3) the six classes of our taxonomy. The features examined are the size (number of words) and element types of the relevant components and the number of different journals in which the relevant information is found.

3.1 Element Size

As in many other information retrieval areas, length normalisation has been used in XML retrieval. Although so far the statistics of element size have only been used across topics, we showed in previous work [8] that there are significant differences in the size of relevant elements between topics. We believe that this type of information could be beneficial for a retrieval system. Figure 1 shows a histogram of the topic sizes for each of the task types.

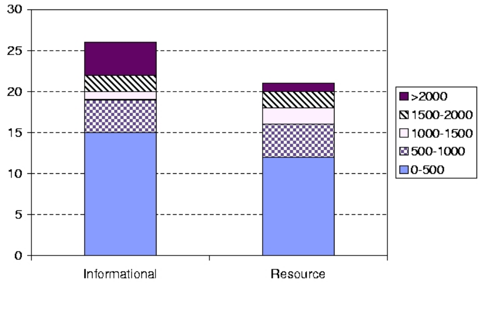

Figure 1: Histogram of topic sizes per task type. The size of a topic is determined by the average of all the sizes of its relevant element.

Although the differences are not very significant, we can see that the *informational* task type tends to require larger elements. The small difference could be due to the fact that our *resource* task includes all information needs where any

[4] st⇒section title, bb⇒citation, fm⇒front matter, fig⇒figure.

type of restriction on the information is made, which does not mean that the required information is *specific* in length (e.g., algorithms or esperimental results). Contrary to our expectation, the differences in relevant sizes for the different degrees of user's collection familiarity and the six classes of our taxonomy, are not significant enough and we can not provide evidence from any relationship.

3.2 Element type

There are more than 150 different element types in the INEX collection. The use of which element types can be relevant for a given task could be an important source of information for the IR system. Table 2 shows the average of the different relevant element types per topic for the different categorizations. As expected, the *informational* task tends to require a wider range of element types. However, we were more surprised by the fact that topics with *high* collection familiarity, the ones that specify collection-specific structural constraints, require also a wider range of element types. A possible explanation of this is that it is difficult to pose the right constraints for a query and, when users are asked to assess all kinds of element types, they find the other types also relevant. This could be considered a type of developement of the user's understanding of their information need when interacting with the documents. However, it could also be the case that some assessors are *too* generous when assessing the relevant information. Further investigation is needed to be able to conclude anything.

Table 2: Average (Median) of the number of different relevant types per topic.

<table>
<tr><td colspan="3"><b>Informational</b>
10.5 (5.5)</td><td colspan="3"><b>Resource</b>
4.6 (4)</td></tr>
<tr><td colspan="2"><b>None</b>
8.8 (4)</td><td colspan="2"><b>Some</b>
3.4 (3)</td><td colspan="2"><b>High</b>
10.5 (7)</td></tr>
<tr><td><b>A</b>
11.1 (8)</td><td><b>B</b>
3.6 (3)</td><td><b>C</b>
30 (24)</td><td><b>D</b>
3.4 (3.5)</td><td><b>E</b>
3.2 (3)</td><td><b>F</b>
6.5 (5.5)</td></tr>
</table>

Regarding the kind of element types relevant to each categorization, we can not see discriminative differences for the different degrees of collection familiarity or for the six classes defined. However, more important differences can be seen for the informational and resource tasks, Figure 2. As expected, *resource* tasks would require more specialized elements such as paragraph or other types, while *informational* tasks would rather require more general elements, such as sections, bodies or articles.

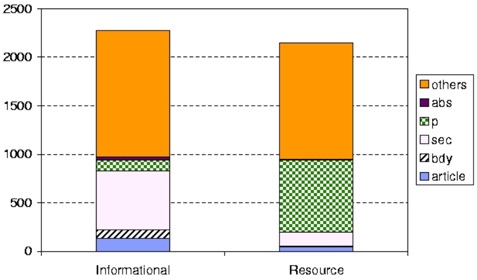

Figure 2: Histogram of element types per task type.

3.3 Journal information

The articles of the INEX collection are clustered into 18 different journals. Our previous work has shown that the use of the information of relevant journals per topic (e.g., on a relevance feedback process) can significantly improve the retrieval effectiveness of the IR system [8]. Table 3 shows the average of the number of different journals containing relevant information for the different categorizations. In general, relevant elements for the *resource* tasks appear in a smaller set of different journals, while relevant information for the *informational* tasks is spread among several journals. Seemingly, this type of information is not affected by the collection familiarity of the user. However, users with high knowledge of the structure seems to require information from a larger set of journals.

Table 3: Average (Median) of the number of different relevant journals per topic

<table>
<tr><td colspan="3"><b>Informational</b>
4.3 (4.5)</td><td colspan="3"><b>Resource</b>
3.1 (2)</td></tr>
<tr><td colspan="2"><b>None</b>
3.6 (3)</td><td colspan="2"><b>Some</b>
2.8 (2)</td><td colspan="2"><b>High</b>
5.2 (4)</td></tr>
<tr><td><b>A</b>
4.3 (4)</td><td><b>B</b>
4 (5)</td><td><b>C</b>
5 (4)</td><td><b>D</b>
2 (2)</td><td><b>E</b>
1.4 (1)</td><td><b>F</b>
5.2 (3.5)</td></tr>
</table>

4. CONCLUSIONS

We have shown in previous work that the use of structural features can be beneficial for retrieval systems [8]. In this paper, we have defined a simple taxonomy of user needs that takes into account the user's familiarity with the collection and we have shown that the distributions of some structural characteristics differ for some of the categories eximined. This suggests that task-specific retrieval techniques should also be considered in XML retrieval.

Unfortunately, the empirical evidence shown in this paper does not imply a significant correlation between user's collection familiarity and the relevance of the different structural features. As future work we plan to further investigate this issue on a larger data set and to refine and extend the taxonomy presented to include other contextual factors such as the user's knowledge on the topic.

5. REFERENCES

[1] N.J. Belkin, R.N. Oddy, and H.M. Brooks. ASK for Information Retrieval: Part I and II. *Journal of Documentation*, 38(2 and 3), 1982.

[2] S. K. Bhavnani, K. Drabenstott, and D. Radev. Towards a Unified Framework of IR Tasks and Strategies. In *ASIST'2001*, pages 340–354.

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

[4] N. Fuhr, N. Gövert, G. Kazai, and M. Lalmas. INEX: INitiative for the Evaluation of XML retrieval. In *SIGIR 2002 Workshop on XML and Information Retrieval*.

[5] M. Lalmas G. Kazai and B.Piwowarski. INEX 2004 Relevance Assessment Guide. In *INEX 2004 Workshop Proceedings*, pages 242–251. Notebook paper.

[6] I. Kang and G. Kim. Query type classification for web document retrieval. In *SIGIR 2003*, pages 64–71.

[7] V. Murdock and W. Croft. Task orientation in question answering. In *SIGIR 2002*, pages 355–356.

[8] G. Ramirez, T. Westerveld, and A.P. de Vries. Structural Features in Content Oriented XML Retrieval. Technical Report INS-E0508, CWI, 2005. `http://db.cwi.nl/rapporten/abstract.php?abstractnr=1864`.

[9] A. Trotman and B. Sigurbjörnsson. Narrowed Extended XPath I (NEXI). In *Lecture Notes in Computer Science*, volume 3493, pages 16–40, 2005.