Structural Features in Content Oriented XML Retrieval

Georgina Ramírez, Thijs Westerveld and Arjen P. de Vries
CWI, Amsterdam, The Netherlands
georgina@cwi.nl, thijs@cwi.nl, arjen@acm.org

ABSTRACT

The structural features of XML components are an extra source of information that should be used in a content-oriented retrieval task on this type of documents. In this paper we explore one of the structural features from the INEX collection [1] that could be used in content-oriented search. We analyse the gain this knowledge could add to the performance of an information retrieval system and present a first approach on how this structural information could be extracted from a relevance feedback process to be used as priors in a language modelling framework.

Categories and Subject Descriptors: H.3.3 Information Storage and Retrieval: Information Search and Retrieval.
General Terms: Algorithms, Experimentation.
Keywords: XML retrieval, relevance feedback, structural features.

1. INTRODUCTION

This paper is part of on going work where we analyse information available in the structure of the documents and show how this information can be useful for content oriented XML retrieval. We analyse the relevance assessments for INEX 2004 [1] and compare the structural information available in the set of elements that has been judged relevant to the structural information in retrieved elements and in the collection in general. The differences in structural characteristics between relevant elements and other elements could be exploited to improve retrieval results.

In this paper, we study the potential of one type of structural information: the containing journal of an element.

2. JOURNAL INFORMATION

The content of the INEX collection consists of eighteen different journals. Each of these journals contains articles discussing a different computer science related fields. Our hypothesis for this type of information is that when a component is assessed relevant for a given topic, the journal where it belongs to will contain elements with a similar content information. This information can be used to increase the a priori belief in relevance of the elements that are contained in that journal.

Table 1 displays general statistics related to journal information. The first row lists statistics regarding the highly relevant components, the second the statistics for all the components assessed with any degree of relevance higher than zero. The number of journals (on average) relevant to a topic, in the most general case, is seven. If we compare this information to the statistics obtained from the results of our retrieval system [2] (third row), we can see that the average number of journals we return per topic is more than twice as high. Even in the best case, the results returned by our system originate from 12 different journals. When we look at this information per topic, even when the number of relevant journals for a topic is very low (2 or 3), the number of different journals returned by our system is very high. A very similar behaviour is observed when analysing the other systems participating at INEX (see Figure 1).

Table 1: General statistics journal: Number of different journals per topic in relevant set and in result set.

Source	Avg	Median	Max	Min
Relevant (highly)	3.6	2	9	0
Relevant (somehow)	7.15	7	16	l2
Results (1500 elements)	16.65	17	18	12

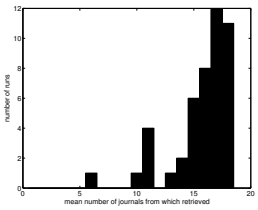

Figure 1: Distribution of the average number of different journals retrieved per run in all INEX runs.

If we look at the distribution of topic terms among the journals, we see that the *journal frequency*, the number of different journals a term occurs in, is very high for most of the topic terms. The topic terms are spread into all the journals and most journals contain more than just a few occurrences of the terms. The *article frequency*, the number of articles containing a term, for these terms in each of the

CIKM'05, October 31–November 5, 2005, Bremen, Germany.
ACM 1-59593-140-6/05/0010.

Table 2: Mean average precision for different ways of using journal information. The plus symbols indicate a significant increase over the baseline using the Wilcoxon signed-rank test at a confidence level of 95% (+) or 99% (++).

baseline	0.0865
filtering optimal	0.1031 (++)
priors full	0.0927 (++)
priors top 20	0.0904
priors top 20 interpolated	0.0918 (+)

journals is also high. Therefore, a typical retrieval system (based on term frequencies in one way or another) will retrieve elements from many different journals even though the relevant elements often appear in only a few journals. This means that the knowledge of the relevant journals per topic could help the retrieval systems to disambiguate these terms and therefore increase its performance. We test this hypothesis experimentally in the next section.

3. REL. FEEDBACK ON STRUCTURE

The main idea of a relevance feedback strategy is to use the knowledge of relevant items to retrieve more relevant items. So far, research has concentrated on using content-related information from the known relevant elements. This section investigates if we can improve retrieval results by using *only* structural information.

In all our experiments we used the Tijah system [2]. We computed the mean average precision (MAP)[1] based on the top 1,500 retrieved elements. Results are compared to a baseline run that uses the basic language model with a linear function of the element length as prior. The MAP for this baseline run is 0.0865. The following subsections discuss using different priors. All results are summarised in table 2.

To investigate the importance of the journal information for a retrieval system, we study the occurrences of journals in the relevant set. For each topic, we order the journals by decreasing number of relevant elements they contain. We then look at the effect of filtering out all elements from the result list for each topic except those belonging to the top N journals for that topic. N is varied from 1 to the total number of relevant journals for the topic. Figure 2 shows the increase in MAP when adding more journals. When only two journals are used to retrieve elements from MAP is already higher than the baseline. The optimal number of journal varies from topic to topic. Using for each topic the optimal number of journals gives an indication of the potential gain from using the journal information. This optimised run has a MAP of 0.1031 (a significant improvement over the baseline).

Instead of only filtering elements it may be useful to reorder elements. To do so, the priors $P(E)$ can be updated and elements that are likely to be relevant will be pushed up in the ranking. Again we look at the full relevance judgements and compute *journal*-priors:

$$P_{journal}(E) = P(rel|journal(E)) \propto \frac{P(journal(E)|rel)}{P(journal(E))}, \quad (1)$$

where $journal(E)$ identifies the journal to which E belongs, $P(journal(E)|relevant)$ is estimated as the fraction of relevant items belonging to the journal and $P(journal(E))$ is the fraction of elements in the collection that belongs to that journal. Note this means that elements that did not appear in the relevant set will get $P_{journal} = 0$ and thus effectively will be removed.

[1]Average over all quantisations.

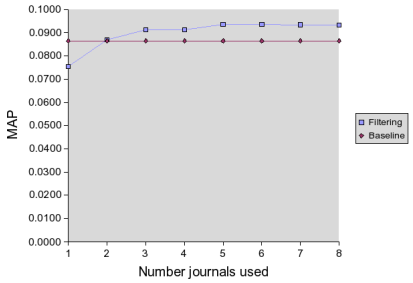

Figure 2: MAP for using increasing number of journals.

Using these journal priors, we obtain a MAP of 0.0927, when we take relevance information from the full assessments, and 0.0904, when we take it from the top 20 elements of the baseline run. Since in the top 20 we may not have seen all journals there is the risk of assigning $P_{journal}(E) = 0$ to elements from journals that do actually contain relevant elements. To avoid this effect of relying too much on what is seen in the top 20, we interpolate $P(journal(E)|rel)$ with the general probability of seeing elements from $journal(E)$. Thus the journal prior becomes:

$$P_{journal}(E) = \frac{\alpha P(journal(E)|rel) + (1-\alpha)P(journal(E))}{P(journal(E))}. \quad (2)$$

With this interpolated prior a small, but significant, improvement over the baseline is obtained, see table 2.

4. DISCUSSION

We have shown that the distribution of a structural characteristic (journal information) differ for relevant elements and other elements. Experiments have shown that the use of this information can improve retrieval effectiveness.

While query terms typically are distributed across many elements in all journals, relevant elements tend to cluster in a few journals. We showed this information is useful in a retrieval setting and leads to significant performance improvements.

We would like to stress that even though the experiments described in this paper are a very naive approach to exploiting the structural information, we improve significantly over the baseline. The experiments reported here do not modify the modelling of content information in any sense. We believe there is great potential for using the information gathered from the structure to improve the modelling of content. For example, to update the background estimates, to recompute IDF values or to do a journal specific query expansion.

5. REFERENCES

[1] 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*, 2002.

[2] V. Mihajlovic, G. Ramirez, A. P. de Vries, , D. Hiemstra, and H. E. Blok. TIJAH at INEX 2004. Modeling Phrases and Relevance Feedback. In *INEX 2004 Workshop Proceedings*, 2004.