Centrum voor Wiskunde en Informatica

REPORT*RAPPORT*

INS

Information Systems

INformation Systems

Structural features in content oriented XML retrieval

G. Ramírez, T.H.W. Westerveld, A.P. de Vries

REPORT INS-E0508 APRIL 2005

CWI is the National Research Institute for Mathematics and Computer Science. It is sponsored by the Netherlands Organization for Scientific Research (NWO).
CWI is a founding member of ERCIM, the European Research Consortium for Informatics and Mathematics.

CWI's research has a theme-oriented structure and is grouped into four clusters. Listed below are the names of the clusters and in parentheses their acronyms.

Probability, Networks and Algorithms (PNA)

Software Engineering (SEN)

Modelling, Analysis and Simulation (MAS)

Information Systems (INS)

Copyright © 2005, Stichting Centrum voor Wiskunde en Informatica
P.O. Box 94079, 1090 GB Amsterdam (NL)
Kruislaan 413, 1098 SJ Amsterdam (NL)
Telephone +31 20 592 9333
Telefax +31 20 592 4199

ISSN 1386-3681

Structural features in content oriented XML retrieval

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. This paper explores three different structural features from the INEX collection 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.

2000 Mathematics Subject Classification: 68P20
1998 ACM Computing Classification System: H.3.3 H.2.8
Keywords and Phrases: XML retrieval; relevance feedback; structural features

Structural Features in Content Oriented XML retrieval

Georgina Ramírez
georgina@cwi.nl

Thijs Westerveld
thijs@cwi.nl

Arjen P. de Vries
arjen@cwi.nl

CWI
P.O. Box 94079, 1090 GB Amsterdam, The Netherlands

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. This paper explores three different structural features from the INEX collection [FGKL02] 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.

1. Introduction

Content-oriented XML retrieval differs from traditional document retrieval, not only in that the retrieval system has to decide which is the most appropriate unit to return to the user, but also because the document contains extra information on how its content is structured. The implicit semantics on how and why the documents are organised in a certain way, might help the information system to retrieve the most relevant information for a user need. The usage of this structural knowledge might not only help to decide what is the best retrieval unit given a query, but it may also help to improve the effectiveness of the content oriented search.

The area has been studied for a number of years now, and an XML retrieval system benchmark (INEX) has been organised in the last three years [FGKL02]. However, so far the structural information in documents has hardly been used, and most systems, including our own, have treated XML retrieval as traditional document retrieval. The main difference is that the retrieved units, traditionally documents, can be any element in the XML tree, ranging from paragraphs and sections to full articles or even complete journals.

This paper analyses the information available in the structure of the documents and shows how this information can be useful. To this end, we analyse the relevance assessments for INEX 2004 [FGKL02] 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. We do not go into much detail on how to obtain relevance information. That process does not differ from the one used in traditional content-based feedback.

The paper is organised as follows. Section 2 gives an overview of work done in the area of content-oriented XML, surface features and relevance feedback on structured documents. Section 3 discusses different features of structural information and analyses the relevance assessments of INEX 2004 for three of them: containing journal, element types and size. In Section 4 we analyse the possible gain the knowledge of this information could add to a retrieval system performance. The section also presents some experiments on using this information in a relevance feedback process. We conclude with a discussion about the results presented in the paper and our main conclusions.

2. Context and Related Work

This section puts the present work in context. We first briefly introduce the INEX benchmark and discuss some of the approaches to XML retrieval that have exploited surface features or structure information in a content-oriented retrieval task. Then we discuss related work on the use of relevance feedback in XML documents and on the use of surface features in general, in various retrieval tasks.

The *Initiative for the Evaluation of XML retrieval* (INEX) [FGKL02] 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. A *highly relevant* component is considered the one assessed as E=3, S=3. More information about this process can be found in [20004]. 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. In this paper we study the importance of structural features for the content-oriented task, where only query terms are given without explicit structural constraints.

In the few years of INEX' existence (2002-2004), a lot of XML retrieval approaches have been presented. We point out here the ones we consider relevant for the paper as they use structural features in the content-oriented retrieval task. As in many other information retrieval areas, length normalisation has been used in XML retrieval [KdRS04, LMR$^+$ar]. These approaches use the length information across topics and ignore individual differences between topics. Another structure feature used often in XML retrieval is the structural relationship between elements to aggregate or propagate scores along the structure of the XML tree [FG01, SKdR04]. These approaches make use of *explicit* structure information to improve the content oriented search and do not make use of individual relevance information. A third group of XML retrieval approaches use the structure to define a subset of possible elements to be retrieved [TO03, MM03]. All these different ways of using structural information treat the structure in a global manner, without considering differences between topics. In this paper, we look at structural characteristics for individual topics, thus opening up the possibilities for relevance feedback on structure.

Different relevance feedback techniques have been used in IR systems. Although extensively used, these techniques have been focusing on the content part of a document. A survey of these techniques applied to different information retrieval models is presented in [RL03]. In [MM04, SKdR04], existing relevance feedback algorithms are applied to query on XML documents. All these approaches are using content-oriented feedback, whereas the feedback studied in this paper is based on structural characteristics. As far as we know, the only work in relevance feedback on structure is done by Mihajlovic et. al [MRdV$^+$04]. They use similar structure features to the ones studied in this paper, but the features are used in an ad hoc manner and no improvement in performance was reported. In this paper, we first analyse the relevance assessments of INEX to study the possible potential of using these type of information in a relevance feedback process and analyse the gain the usage of this information would add to the retrieval systems performance.

Outside the area of XML retrieval, surface features, i.e., anything other than content information, have been studied mainly in the context of web retrieval. A host of work exists that studies ways to exploit the hyperlink structure between documents. See for example [BP98, Kle99, BH98]. Kraaij et al. [KWH02] demonstrate that using information obtained from URL-length can improve performance when querying for homepages. Also in the HARD track at TREC [All03], surface features are studied, but there the features describe characteristics of the searcher rather than the documents. In other information retrieval areas the only surface feature that has been used widely is document length, which is typically used for normalisation.

3. Structural information

This section identifies different aspects of the structure of documents that could give important information to a retrieval system, and analyses three of them in depth.

The purpose of structuring documents is to organise their content and to divide it into smaller, semantically similar, units. The markup of this documents, used to organise the content of the documents, can provide different types of information. As an example, take a look at the XML fragment in Figure 3, extracted from the INEX collection. The markup tag **p** does not give much information in itself. However, with some knowledge of the collection, this tag might provide to the retrieval system extra information like the average size of this type of elements, the kind of content they contain, their role in the hierarchy or their location within the structure of the document (e.g., leaf nodes). Apart from all this *implicit* information, sometimes the markup explicitly gives information about the content or style of the text. This would be the case of tags as **author** or **italics** respectively. This other type of information can again be used by the retrieval system, which could use it to re-weight terms appearing in specific element types, for instance, in titles.

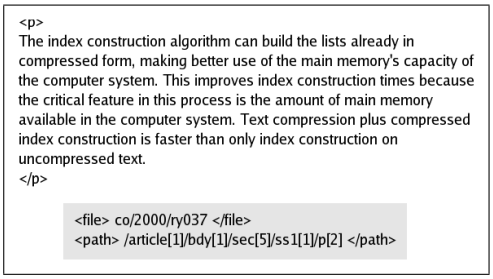

Figure 1: XML fragment extracted from the INEX collection.

If we now look at the grey area in Figure 3, we see yet another type of information: the information to locate this XML fragment within the INEX collection. The **file** tag is used to locate the article and the **path** indicates the structural path within the article. The structural information to locate the article provides some extra information about the organisation of the files and therefore about the content of this fragment: the first part of the path (**co**) indicates the journal it belongs to (**Computer**), and the second part (**2000**) is used to indicate the year of publication of the article. This information could obviously help a search: for example, articles about *content based music retrieval* are more likely to occur in *Multimedia* or *Transactions on Pattern Analysis and Machine Intelligence* than in *Transactions on Parallel & Distributed Systems*. The structural information about the location of the fragment within the article provides again extra information as the level where the fragment is located in the document hierarchy or the context where this element appears. Finally, the *size* of the text contained in the fragment might provide information about the type of element.

We believe that all structural information contained in the XML fragments can help the information retrieval systems to refine their content search and to decide which is the best retrieval unit to return to the user. The assumption we follow is that the structure exists for a reason and tells something about the document. Therefore, the structural information is discriminative and should be used for retrieval purposes.

To study the potential of this type of information for content-oriented XML retrieval, we analyse different aspects of the relevance judgements for INEX 2004. We study three different structural features from the XML components: the containing journal of an element, the element type and the element size.

Table 1: List of journals included in the INEX collection.

an	IEEE Annals of the History of Computing
cg	IEEE Computer Graphics and Applications
co	Computer
cs	Computing in Science & Engineering
dt	IEEE Design & Test of Computers
ex	IEEE Intelligent Systems
ic	IEEE Internet Computing
it	IT Professional
mi	IEEE Micro
mu	IEEE Multimedia
pd	IEEE Parallel & Distributed Technology
so	IEEE Software
tc	IEEE Transactions on Computers
td	IEEE Transactions on Parallel & Distributed Systems
tg	IEEE Transactions on Visualization and Computer Graphics
tk	IEEE Transactions on Knowledge and Data Engineering
tp	IEEE Transactions on Pattern Analysis and Machine Intelligence
ts	IEEE Transactions of Software Engineering

Table 2: General statistics journal: Number of different journals per topic in relevant set and in result set.

Source	**Avg**	**Median**	**Max**	**Min**
Relevant (3E3S)	3.6	2	9	0
Relevant (all)	7.15	7	16	2
Results (1500 elements)	16.65	17	18	12

3.1 Journal

The content of the INEX collection consists of eighteen different journals. Each of these journals contains articles discussing a different computer science related field. The journals included in the INEX collection and their abbreviations are listed in table 1. 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.

This subsection analyses if, according to the relevance assessments, the use of this clustering information could improve a content oriented search.

Table 2 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 [MRdV$^+$04] (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. The first observation we can obtain from this statistics is that the knowledge of the relevant journals given a topic should improve our results considerably. Figure 2 presents this information per topic. Note that, even when the number of relevant journals for

a topic is very low (e.g. topics 162 or 168), the number of different journals returned by our system is very high.

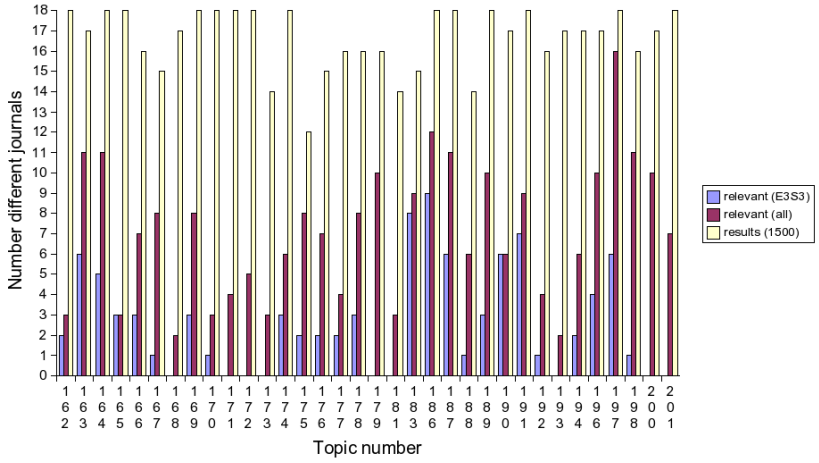

Figure 2: Number of different journals per topic; relevant set vs. result set.

We then investigated if the other systems participating in INEX return elements from a comparable number of different journals. Figure 3, shows the distribution of the average number of journals retrieved per run. We can see, our system's pattern is followed by most of the runs.

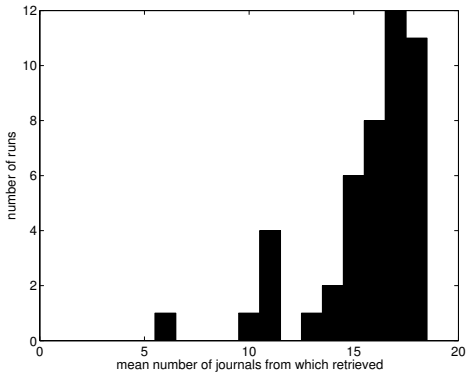

Figure 3: Distribution of the average number of different journals retrieved per run in all INEX runs.

If we look at the distribution of topic terms among the journals (Figure 4) 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, as Figure 5 shows, 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 journals is also high. Analysing the terms for a specific topic shows the same behaviour. Figure 6 shows the *article frequencies* of the topic terms in topic number 173 (content based music retrieval) in the different journals.

The distribution of term counts shows that 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 section 4.

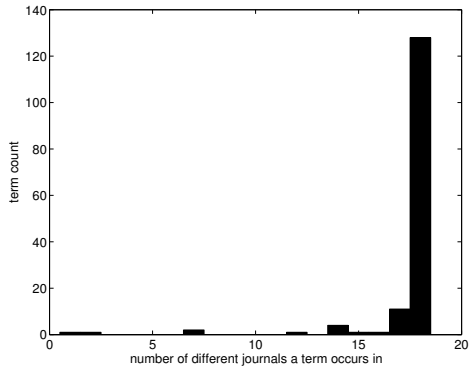

Figure 4: Journal frequency of the topic terms.

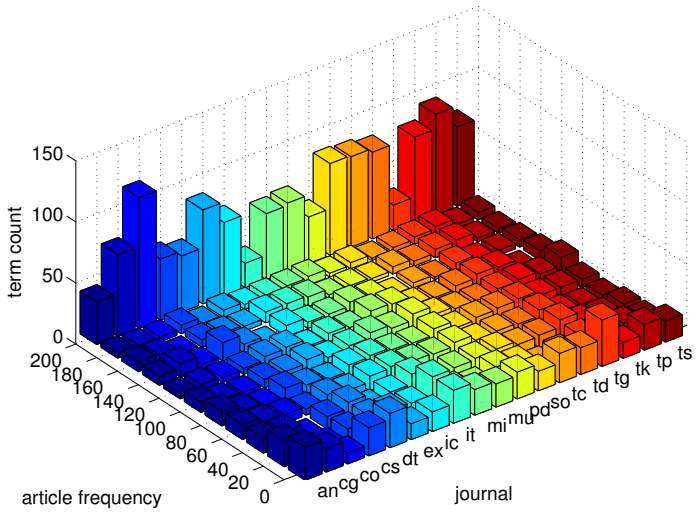

Figure 5: Article frequency of topic terms per journal. General.

3.2 Element type

The INEX collection contains more than 150 different element types. Although some of these are not very sensible retrieval units (e.g., style components), the diversity remains large. This subsection analyses if knowledge of the relevance of the different types of elements can help to improve a content oriented search.

Table 3 presents general statistics related to element type information. The first row shows statistics regarding the highly relevant components,the second statistics for all components assessed with any degree of relevance higher than zero. The number of different element types relevant for a topic is relatively large. On average, twenty-two different types of elements are relevant for a topic and therefore, possible retrieval units for the search systems. Comparing this information to the statistics obtained from the results of our retrieval system (third row in Table 3), we can see that the average number of element types we return per topic is too high, although the difference is not as large as in the journal case. That is because our retrieval model already uses length normalisation on the size of the elements to return, pushing very small elements down the ranked list. Still, looking at the statistics, we predict that knowledge of relevant element types given a topic could improve results. Figure 7 presents this information per topic. Note that, once more, even when the number of relevant

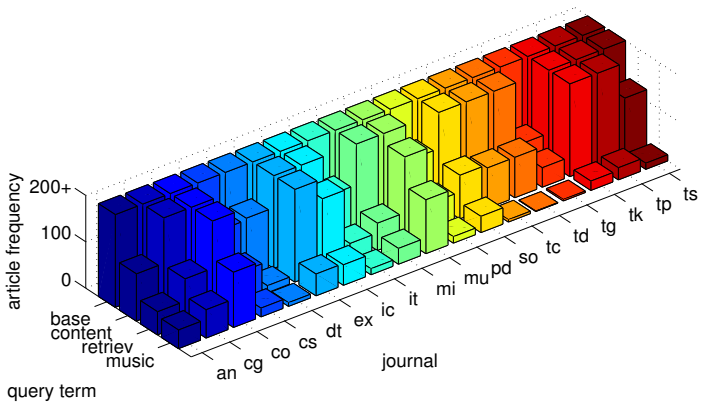

Figure 6: Article frequency of terms per journal for topic 173.

Table 3: General statistic elements: Number of element types per topic in relevant set and result set.

Source	Avg	Median	Max	Min
Relevant (3E3S)	8.6	4	31	0
Relevant (all)	22.32	19	60	6
Results (1500 elements)	35.03	35	51	12

elements for a topic is very low (e.g., topics 165 or 190), the number of different elements returned by our system is very high.

Like we did in the journals case, we analysed the behaviour of all INEX runs. Figure 8 shows the distribution of the average number of element types returned. We can see that our run is not representative, as more than the third of the runs returned a small number of different element types. This is because some of the systems use (small) pre-defined subsets of element types to use as the only possible retrieval units. Some runs are even restricted to a single element type.

Although the differences are not as clear as in the journal case, it seems that the knowledge of the type of elements that are relevant or preferred by a user, could help to improve system performance both for runs that return many different element types and for those that restrict their results to a subset of types. We study this hypothesis in Section 4.

3.3 Size

The size of elements is known to be an important factor in the prior probability of the element being relevant to any topic. Larger elements are more likely to be relevant. This is the case in traditional document retrieval, but also in XML retrieval [KdRS04]. So far, the statistics of element size have only been used across topics, the individual differences between topics have been mostly ignored.

Analysing the element sizes in the set of relevant elements (for all topics) and in the collection, we find the well-known distributions: the collection contains many small elements, but the relevant elements tend to be larger, see Figure 9. Looking at the size distributions of relevant elements for individual topics, we find most topics follow the general trend. However, topics with a different behaviour exist. Some topics tend to have smaller elements in their relevant set, others prefer very large elements. Figure 10 shows an example of each case.

Differentiation between topics may lead to a greater improvement in retrieval effectiveness than treating all topics equally. We study this effects in next section.

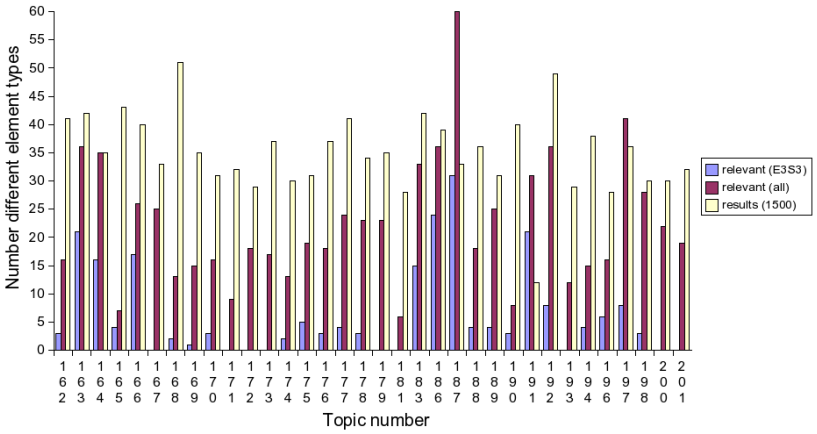

Figure 7: Number of different element types per topic; relevant set vs. result set.

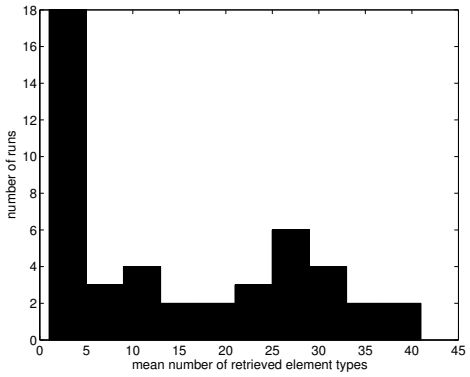

Figure 8: Distribution of the average number of different element types retrieved per run in all INEX runs.

4. Relevance Feedback on Structure

The main idea of 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. To this end we exploit the differences in characteristics between relevant elements and non-relevant elements as identified in the previous section. Obtaining these characteristics is a hard problem in itself and is not addressed extensively here. We mainly test if knowing something about the structural characteristics of the elements wanted by the user could improve retrieval effectiveness. To this end, we do a retrospective analysis in which we take the full relevance judgements and incorporate the derived statistics in our retrieval model. Of course, using the knowledge obtained from the full relevance judgements is not realistic. To test whether the use of structural information has potential in a practical setting, we also experiment with obtaining this information from relevant elements that are retrieved in the top 20 of our baseline run. This setting mimics the situation of a user providing feedback on the top 20 documents.

4.1 Updating priors in a language modelling framework

Our experiments use a system based on simple statistical language models [PC98, Hie98, MLS99]. Using Bayes' rule, the probability of a element E given a query Q can be estimated as the product

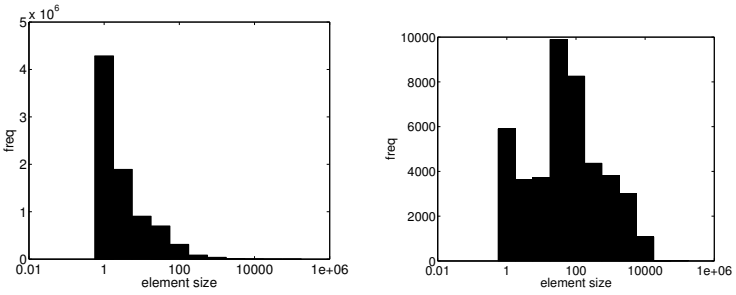

Figure 9: Distribution of element sizes in the collection and in relevant elements.

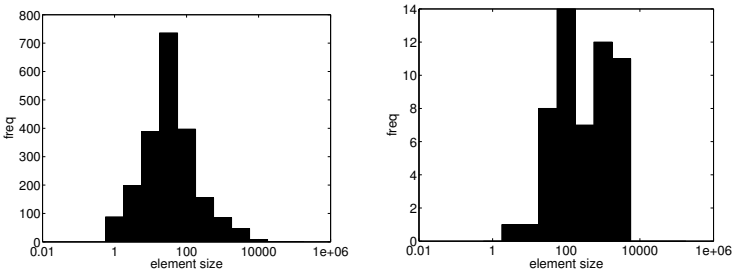

Figure 10: Distributions of element sizes in relevant elements for individual topics.

of the probability of generating the query terms q_i from the element's language model and the prior probability of the element:

$$P(E|Q) \propto \prod_{q_i \in Q} P(q_i|E)P(E) \tag{4.1}$$

Typically, little prior knowledge about the probability of an element is available, and uniform priors are used or longer elements are assumed to be more likely to contain relevant information and $P(E)$ is taken to be related to the element's length (cf. [KdRS04]). However, once we have some information about relevant elements, for example from a user's relevance judgements, we can use this information to update the priors. From the judgements, we can discover the characteristics of relevant elements and update the priors in such a way that elements with similar characteristics are favoured.[1] Note that this does not require updating of the content models, i.e., the elements' language models do not change.

In addition to this principled way of incorporating priors in the language modelling framework, we look at the effect of filtering out all elements with characteristics that do not occur in the relevant set for a given topic. We also experiment with keeping only those elements that have the most frequent characteristics in the relevance set. This gives a first indication of the contribution of a specific structural characteristic.

All experiments are compared to a baseline run that uses the basic language model with a linear function of the element length as prior. The mean average precision (MAP) for this baseline run is 0.0865. The MAPs for all experiments are summarised in table 4.

Obviously, the information from relevance judgements can also be used in the traditional fashion to improve the content modelling, or to improve or extend the original query. Section 5 discusses the

[1] Strictly speaking $P(E)$ can no longer be called a prior, since it depends on the topic at hand.

Table 4: Mean average precision for different ways of using structural information (journal, element types and element size). The plus symbols indicate a significant increase over the baseline using the Wilcoxon signed-rank test at a confidence level of 95% (+) or 99% (++).

	journal	e.type	e.size
baseline	0.0865	0.0865	0.0865
filtering optimal	0.1031 (++)	0.0960 (++)	-
priors full	0.0927 (++)	0.0943	0.0892
priors top 20	0.0904	0.0791	-
priors top 20 interpolated	0.0918 (+)	0.0820	-

interplay between the two ways of using feedback.

4.2 Journal

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 11 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 according to Wilcoxon's signed-rank test[2]). Figure 12 shows the average precision per topic for the baseline run (results) and this optimized run.

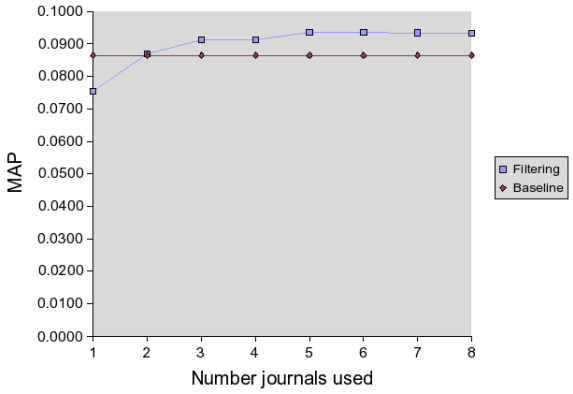

Figure 11: MAP for using increasing number of journals.

Instead of only filtering elements it may be useful to re-order 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:

$$P_{journal}(E) = P(rel|journal(E)) \propto \frac{P(journal(E)|rel)}{P(journal(E))}, \tag{4.2}$$

[2]Throughout this paper significance is tested using the Wilcoxon signed-rank test at a 95% confidence level (unless stated otherwise).

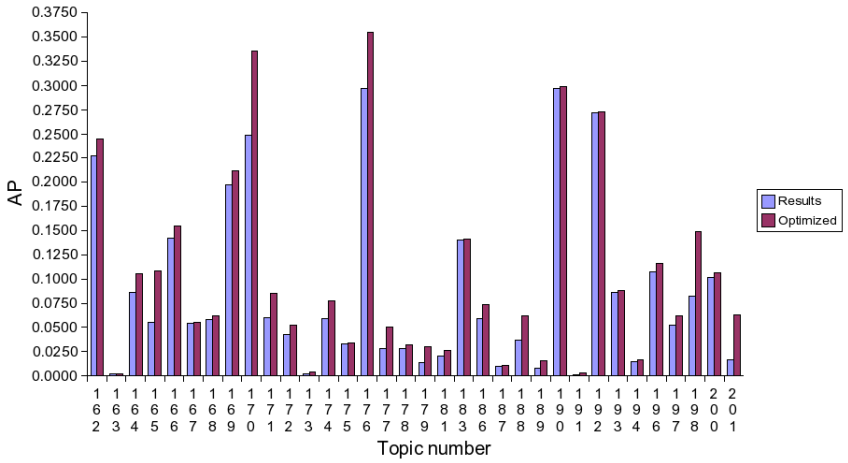

Figure 12: Average precison per topic; baseline vs. optimal journal filtering.

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.

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:

$$P_{journal}(E) = \frac{\alpha P(journal(E)|rel) + (1-\alpha)P(journal(E))}{P(journal(E))}. \tag{4.3}$$

With this interpolated prior a small, but significant, improvement over the baseline is obtained, see table 4.

4.3 Element type

For element type, we performed a similar retrospective analysis of filtering element types that did not occur in the relevant set and of updating priors based on relevance information. The element priors are defined as:

$$P_{element}(E) = P(rel|element(E)) \propto \frac{P(element(E)|rel)}{P(element(E))}, \tag{4.4}$$

where $element(E)$ identifies the element type of E and the probabilities are estimated like in the journal case. Also, an adapted prior is tested to cater for element types that are not observed in the top 20 (cf. Eq. 4.3). Again, we find improvements over the baseline, but in this case only the filtering run shows a significant improvement. Figure 13 shows the change in MAP as more and more elements types are allowed in the result set; Figure 14 shows the results at the individual topic level for the filtering run in which we took the optimal number of element types for each topic. The optimal number of element types varies from 2 to 60; for some topics the best score is obtained when using all element types, i.e., without any filtering. This contrasts the approaches at INEX of retrieving only a pre-defined set of element types. The types typically used in those approaches (e.g., paragraph and

section) are found to be important in our optimal runs, but removing other types would harm results.

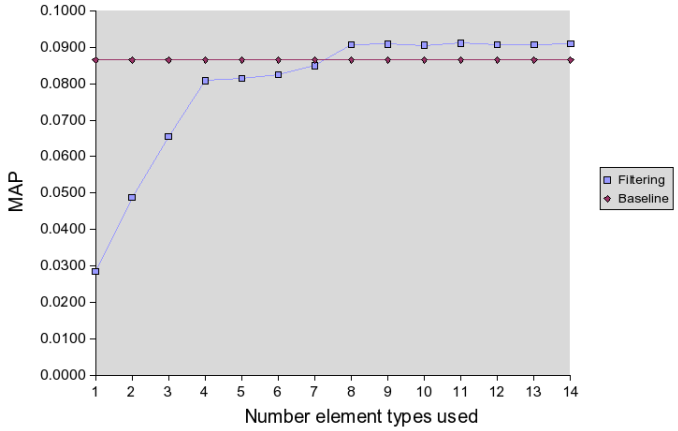

Figure 13: MAP for using increasing number of element types.

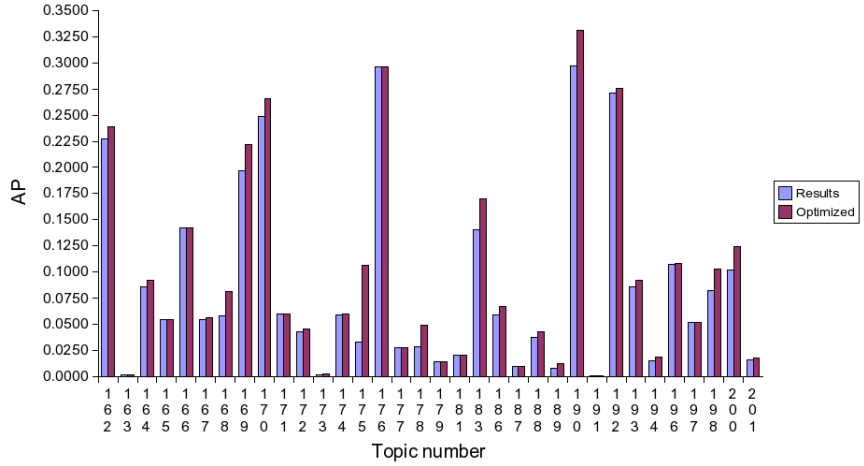

Figure 14: Average precison per topic; baseline vs. optimal element filtering.

4.4 Element size

For element size there is probably not enough information in the top 20 retrieved elements to get accurate estimates. Therefore, for this type of information, we only experiment with taking information from the full set of relevance judgements. Also, most topics have relevant elements of all different sizes (see Figure 10), thus filtering based on size would have no effect. Therefore, we only look at the effects of using topic specific element size priors.

$$P_{size}(E) = P(rel|size(E)) \propto \frac{P(size(E)|rel)}{P(size(E))}. \tag{4.5}$$

$P(size(E)|rel)$ and $P(size(E))$ are estimates on the frequencies of elements of a given size in the set of relevant elements and in the collection. Element sizes are binned into 11 bins on a log scale, ranging from elements with a single term to elements with over 50,000 terms (cf. Figure 10). Based on the

element frequencies in each bin, we obtain a size prior for each bin. Using this prior, the resulting MAP is 0.0892, effectively the same as the baseline score. A possible explanation for this lack of improvement is the fact that the baseline already contains a prior that is based on size. Apparently, a combination of topic specific prior and basic prior does not give an improvement. When the topic specific prior is used on a content only run (i.e., a run without any priors), we reach a MAP of 0.0675, a significant improvement over the content only baseline (0.0492), but significantly worse than the generic size prior. An explanation for the superiority of the generic prior could be the fact that the generic prior is a function of the length, and thus has a much finer granularity then the broad bins used for the topic specific ones. However, using smaller bins would increase the likelihood of inaccurate estimates. An alternative would be to fit a functional form to the empirical priors obtained from the relevance judgements, but this would probably annul the differences between topics.

5. Discussion

We have showed that the distributions of a number of structural characteristics differ for relevant elements and other elements. This means that this information can be useful if we can get our hands on it. The structural information can be elegantly incorporated in the language modelling framework by updating the element priors. Experiments showed that indeed using some of these features can improve retrieval effectiveness.

Especially the information of journals that are likely to contain relevant information is an important clue. 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.

The information obtained from relevant element types has not led to a significant gain in retrieval effectiveness so far. One explanation for this is that the baseline already contains a size prior and therefore naturally prefers elements of a certain size. Since element type is related to size (e.g., paragraphs are typically larger than titles) the addition of an element type prior does not help much. In fact, when the element type prior is used on the content model without size prior, it gives a significant improvement. The interplay between the two priors needs further study. Another explanation for the lack of success of the element type prior is the large number of different elements existing in the collection. Future research has to show whether grouping element types into clusters of similar types (e.g., paragraph, section) would yield more accurate estimates and improved results.

Apparently the sizes of relevant elements do not differ much from one topic to the next, and the use of a generic size prior for all topics performs as least as good as a topic specific size prior.

In this work, we only studied structural features in the INEX collection. Nevertheless, the ideas presented here can be used in any collection of XML material. The specific features that found to be useful may differ, but we believe that the same principles apply. At least an analysis of surface features like the one carried out in this paper could be done to identify features of possible use.

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 the knowledge about journals that are likely to contain relevant information could be used to update the background estimates, or recompute IDF values. This way, the system will focus on terms that are distinguishing within the relevant journals rather than in the whole collection. Also, the journal information allows a system to do a journal specific query expansion and run separate expanded queries against promising journals.

We conclude that information based on structural characteristics of (relevant) elements should be exploited in XML retrieval. First of all, because this is the key feature that really distinguishes the task from traditional document retrieval; but, moreover, because it is a valuable source of information that can enhance the modelling of both content and structure and thus improve retrieval effectiveness.

References

[20004] INEX 2004. INEX 2004 Relevance Assessment Guide. In Norbert Fuhr, Saadia Malik, and Mounia Lalmas, editors, *INEX 2004 Workshop Proceedings*, 2004. notebook paper.

[All03] James Allan. HARD track overview in TREC 2003; high accuracy retrieval from documents. In *The Twelfth Text Retrieval Conference*, 2003.

[BH98] K. Bharat and M. R. Henzinger. Improved algorithms for topic distillation in a hyperlinked environment. In *Proceedings of the 21st Annual International ACM SIGIR C onference on Research and Development in Information Retrieval*, pages 104–111, 1998.

[BP98] Sergey Brin and Lawrence Page. The anatomy of a large-scale hypertextual Web search engine. *Computer Networks and ISDN Systems*, 30(1–7):107–117, 1998.

[FG01] N. Fuhr and K. Großjohann. XIRQL: A query language for information retrieval in XML documents. In W. B. Croft, D. Harper, D. H. Kraft, and J. Zobel , editors, *Proceedings of the 24th Annual International Conference on Research and development in Information Retrieval*, pages 172–180. ACM, 2001.

[FGKL02] Norbert Fuhr, Norbert Gövert, Gabriella Kazai, and Mounia Lalmas. INEX: INitiative for the Evaluation of XML retrieval. In *Proceedings of the SIGIR 2002 Workshop on XML and Information Retrieval*, 2002. `http://www.is.informatik.uni-duisburg.de/bib/xml/Fuhr_etal_02a.html`.

[Hie98] Djoerd Hiemstra. A linguistically motivated probabilistic model of information retrieval. In Christos Nicolaou and Constantine Stephanidis, editors, *Proceedings of the Second European Conference on Research and Advanced Technology for Digital Libraries (ECDL)*, volume 513 of *Lecture Notes in Computer Science*, pages 569–584. Springer-Verlag, 1998.

[KdRS04] Jaap Kamps, Maarten de Rijke, and Börkur Sigurbjörnsson. Length normalization in xml retrieval. In *SIGIR '04: Proceedings of the 27th annual international conference on Research and development in information retrieval*, pages 80–87. ACM Press, 2004.

[Kle99] Jon M. Kleinberg. Authoritative sources in a hyperlinked environment. *Journal of the ACM*, 46(5):604–632, 1999.

[KWH02] Wessel Kraaij, Thijs Westerveld, and Djoerd Hiemstra. The importance of prior probabilities for entry page search. In *Proceedings of the 25th Annual International ACM SIGIR C onference on Research and Development in Information Retrieval*, pages 27–34. ACM Press, 2002.

[LMR$^+$ar] Johan List, Vojkan Mihajlovic, Georgina Ramirez, Arjen P. de Vries, Djoerd Hiemstra, and Henk Ernst Blok. TIJAH: Embracing IR Methods in XML Databases. *Information Retrieval Journal*, to appear.

[MLS99] D. R. H. Miller, T. Leek, and R. M. Schwartz. A hidden Markov model information retrieval system. In *Proceedings of the 22nd Annual International ACM SIGIR C onference on Research and Development in Information Retrieval*, pages 214–221, 1999.

[MM03] Yosi Mass and Matan Mandekbrod. Retrieving the most relevant XML Components. In Norbert Fuhr, Saadia Malik, and Mounia Lalmas, editors, *INEX 2003 Workshop Proceedings*, 2003. `http://inex.is.informatik.uni-duisburg.de:2003/proceedings.pdf`.

[MM04] Yosi Mass and Matan Mandelbrod. Relevance Feedback for XML Retrieval. In Norbert Fuhr, Saadia Malik, and Mounia Lalmas, editors, *INEX 2004 Workshop Proceedings*, 2004. notebook paper.

[MRdV+04] Vojkan Mihajlovic, Georgina Ramirez, Arjen P. de Vries, , Djoerd Hiemstra, and Henk Ernst Blok. TIJAH at INEX 2004. Modeling Phrases and Relevance Feedback. In Norbert Fuhr, Saadia Malik, and Mounia Lalmas, editors, *INEX 2004 Workshop Proceedings*, 2004. notebook paper.

[PC98] J. M. Ponte and W. Bruce Croft. A language modeling approach to information retrieval. In *Proceedings of the 21st Annual International ACM SIGIR C onference on Research and Development in Information Retrieval*, pages 275–281, 1998.

[RL03] Ian Ruthven and Mounia Lalmas. A survey on the use of relevance feedback for information access systems. *Knowl. Eng. Rev.*, 18(2):95–145, 2003.

[SKdR04] B. Sigurbjörnsson, J. Kamps, and M. de Rijke. An element-based approach to XML retrieval. In N. Fuhr, M. Lalmas, and S. Malik, editors, *Proceedings of the Second Workshop of the INitiative for the Evaluation of XML retrieval (INEX)*, ERCIM Publications, 2004.

[TO03] Andrew Trotman and Richard A. O'Keefe. Identifying and Ranking Relevant Document Elements. In Norbert Fuhr, Saadia Malik, and Mounia Lalmas, editors, *INEX 2003 Workshop Proceedings*, 2003. `http://inex.is.informatik.uni-duisburg.de:2003/proceedings.pdf`.