

A survey on XML focussed component retrieval

Karen Pinel-Sauvagnat & Mohand Boughanem

IRIT-SIG/RFI

118 route de Narbonne

31 062 Toulouse Cedex 4

{sauvagnat, bougha}@irit.fr

Abstract

Focussed XML component retrieval is one of the most important challenge in the XML IR field. The aim of the focussed retrieval strategy is to find the most exhaustive and specific element in a path, i.e. to retrieve elements that focus on the user need, without nested elements. In this paper, we introduce a relevance propagation method dealing with focussed XML component retrieval. Many experiments are carried out with the INEX 2005 test suite to define what are the main characteristics of relevant elements in focussed retrieval and to compare such characteristics with those of relevant elements in thorough retrieval (where the aim is to find all relevant elements in the collection). Our main findings are the following. First, a term weighting scheme taking into account the importance of terms in elements and both in collection of elements and collection of documents is useful. Moreover, the introduction of component length as a threshold on results or used in a weighted propagation function improves significantly the results. Third, contextual relevance seems not to be useful, which contradicts results obtained by state-of-the-art methods for non-focussed retrieval. At last, the use of structural hints increases up to 50% performances we obtained when using queries composed only of simple keyword terms.

1. Introduction

The growing number of XML documents leads to the need for appropriate retrieval methods which are able to exploit the specific features of this type of documents. Hierarchical document structure can be used to return specific document components instead of whole documents to users.

These last years, many methods were proposed in the literature for selecting relevant elements. Some of these methods are based on the vector space model (Grabs and Scheck, 2002; Mass and Mandelbrod 2004; Kakade and Raghavan, 2005), or on the probabilistic model, either using a relevance propagation method (Fuhr and Grossjohann, 2004) or using language models (Kamps et al., 2004). Such methods have been evaluated since 2002 by means of the INEX (*Initiative for the Evaluation of XML Retrieval*) campaign. This initiative provides an opportunity for participants to evaluate their XML retrieval methods using uniform scoring procedures and a forum for participating organizations to compare their results.

However, a major problem still remains until 2005: the overlap of results elements, i.e. results elements which are nested within each others. From the end-user perspective, overlap is not a desirable property since the same information may be returned multiple times. For example, a user may not be satisfied when seeing in the same result list a *section* element and a *subsection* element contained in the previous *section* element. Until 2005, official metrics were based on the traditional evaluation models and overlap elements were necessary to obtain good results. This can be explained as follows. Two dimensions of relevance are used by the participants when doing the assessments: *exhaustivity* (e) and *specificity* (s). Exhaustivity is defined as a measure of how exhaustively an element discusses the topic of request, while specificity is defined as a measure of how focused the element is on the topic of request. For both dimensions, a graded-scale of values is used. Inference rules implies that when a node is judged relevant, its

parent should also be judged relevant: it can be less specific, but its exhaustivity is always superior or equal. Thus, the hierarchical structure of XML documents results in a recall-base consisting of overlapping reference components. As a result, perfect recall according traditional metrics used in INEX (-metrics are based on recall-precision-) can only be reached by systems that return all the reference components in the recall-base, including all the overlapping elements (Kazai et al. 2004). Best ranked approaches in INEX 2003 and 2004 had for example nearly 80% of nodes overlap.

The XML Cumulated Gain Metric proposed in (Kazai et al. 2004) aims at solving this problem, and was used as official metric in the INEX 2005 campaign. As a consequence, a task consisting of focussed retrieval and forbidding nested elements was at last introduced in the campaign. The aim of the focussed retrieval strategy is to find the most exhaustive and specific element in a path, i.e. to retrieve elements that focus on the user need, without any overlapping elements. Before 2005, only a few approaches have proposed focussed XML retrieval strategies (Kekalainen et al., 2004; Zwolf et al., 2004). They were however not or badly evaluated, since no appropriate evaluation metric was available. *Our aim in the paper is not to present a new approach for XML retrieval, but to validate (or refute) conclusions on relevant element properties obtained by state-of-the art methods on non-focussed retrieval (also called thorough retrieval in the INEX evaluation campaign).*

According to state-of-the-art methods on thorough retrieval and to experiments with our own search method, using a thorough retrieval strategy leads to the following conclusion on relevant elements:

- concerning the term weighting scheme, the importance of terms in elements as well as in the whole collection is necessary (Sauvagnat et al., 2005; Trotman, 2005). Component length can also be introduced when weighting terms (Kamps et al. , 2004).
- component length should not be used as a threshold on results but can be used to emphasize the importance of small descendant nodes when evaluating nodes relevance score (Sauvagnat et al., 2005).
- contextual relevance (i.e. relevance of documents) is very useful (Mass and Mandelbrod, 2005; Arvola et al., 2005)
- the use of structural hints is at least useful to enhance performance at low recall levels (Kamps et al., 2005; Sauvagnat et al., 2006).

In this paper, we introduce a relevance propagation method dealing with focussed XML component retrieval. Many experiments are carried out with the INEX 2005 test suite on content-only queries for defining what are the main characteristics of relevant elements in focussed retrieval. These characteristics are compared to previous experiments obtained with traditional metrics. We carried out experiments concerning:

- (i) the term weighting scheme,
- (ii) component length,
- (iii) contextual relevance,
- (iv) the use of structural hints.

The rest of the paper is organized as follows. Section 2 presents our baseline model, which uses a relevance propagation function. The INEX 2005 test suite and the associated metrics are described in section 3. Section 4 presents experiments on term weighting schemes, section 5 experiments on component length, section 6 experiments on contextual relevance and section 7 experiments on the use of structural hints.

2. Baseline model

We consider that a structured document sd_i is a tree, composed of simple nodes n_{ij} , leaf nodes ln_{ij} and attributes a_{ij} . Leaf nodes ln_{ij} are content-bearer whereas other nodes only give indication on structure.

During content-only query processing, relevance values are assigned to leaf nodes and relevance scores of inner nodes are then computed dynamically.

2.1 Evaluation of leaf nodes weights.

The first step in query processing is to evaluate the relevance value of leaf nodes ln according to the query. Let $q=\{t_1, \dots, t_n\}$ be a query composed of simple keyword terms (also called Content-Only (CO) query). Relevance values are computed thanks to a similarity function $RSV(q, ln)$.

$$RSV(q, ln) = \sum_{i=1}^n w_i^q \times w_i^{ln} \quad [1]$$

Where: w_i^q is the weight of term i in query q and w_i^{ln} is the weight of term i in leaf node ln .

2.2 Relevance propagation.

In our model, each node in the document tree is assigned a relevance value which is function of the relevance values of the leaf nodes it contains. Terms that occur close to the root of a given subtree seem to be more significant for the root element than ones on deeper levels of the subtrees. It seems therefore intuitive that the larger the distance of a node from its ancestor is, the less it contributes to the relevance of its ancestor. This affirmation is modeled in our propagation formula by the use of the $dist(n, ln_k)$ parameter, which is the distance between node n and leaf node ln_k in the document tree, i.e. the number of arcs that are necessary to join n and ln_k . Moreover, it is also intuitive that the more relevant leaf nodes a node has, the more relevant it is. We then introduce in the propagation function the $|L_n^r|$ parameter, which is the number of n descendant leaf nodes having a non-zero score. The relevance value r_n of a node n is finally computed according to the following formula:

$$r_n = \frac{1}{|L_n^r|} \sum_{i=1..N} \alpha^{dist(n, ln_k)-1} \times RSV(q, ln_k) \quad [2]$$

where ln_k are leaf nodes being descendant of n and N is the total number of leaf nodes being descendant of n .

Then, in order to remove nodes overlap, we use the following strategy: for each relevant path, we keep the most relevant node in the path. The results set is then parsed again, to eliminate any possible overlap among ideal components.

2.3 Discussion

Many relevance propagation methods can be found in the literature (Abolhassani and Fuhr, 2004; Anh and Moffat, 2002; Grabs and Scheck, 2002; Gövert et al., 2002). Our approach differs from these previous works on two main points. The first point is that all leaf nodes are indexed,

because we think that even the smallest leaf nodes can be relevant or can give information on the relevance of its ancestors. Advantages of such an approach are twofold: first, the index process can be done automatically, without any human intervention and the system will be so able to handle heterogeneous collections automatically; and secondly, even the most specific query concerning the document structure will be processed (in case of Content and Structure queries), since all the document structure is kept.

The second point is that the propagation is made step by step and takes into account the distance that separate nodes in the document tree.

Our aim here is not to present a new propagation method, but to clearly identify what are the main characteristics of relevant elements according to the new evaluation metrics.

3. The INEX 2005 evaluation campaign

3.1 Collection and topics.

We used the well-known INEX framework to evaluate our focussed retrieval strategy. The 2005 test collection completes the one used during the last years and is composed of more than 17000 documents with extensive XML markup, extracted from IEEE Computer Society journals published between 1995 and 2004.

Experiments presented here are related to the Content-Only (CO) task. Queries with content-only conditions are requests that ignore the document structure and contain only content related conditions, e.g. only specify what an element should be about without specifying what that component is. The 2005 CO task is composed of 29 topics and of the associated relevance judgments.

Relevance judgments for each query are done by the participants. Two dimensions of relevance are used: exhaustivity (e) and specificity (s). Exhaustivity is measured using a 4-level scale: highly exhaustive (e=2), somewhat exhaustive (e=1), not exhaustive (e=0), too small (e=?). Specificity is measured on a continuous scale with values in [0,1], where s=1 represents a fully specific component (i.e. one that contains only relevant information).

3.2 Metrics

Relevance metrics used in 2005 are different from those used in the previous years and are based on XCG and ep/gr metrics (Kazai and Lalmas; 2005). In order to obtain evaluation with these two metrics, the two dimensions of relevance (exhaustivity and specificity) need to be quantised into a single relevance value. Quantisation functions for 2 user standpoints are used:

- a strict quantisation to evaluate whether a given retrieval approach is able of retrieving highly exhaustive and highly specific document components:

$$f_{strict}(e, s) = \begin{cases} 1 & \text{if } e = 2 \text{ and } s = 1 \\ 0 & \text{otherwise} \end{cases} \quad [3]$$

- a generalised quantisation has been used in order to credit document components according to their degree of relevance :

$$f_{generalised}(e, s) = e \times s \quad [4]$$

An ideal recall-base is defined to evaluate the CO-Focussed task. An ideal recall-base is a subset of the full recall-base, where overlap between reference elements is removed in order to form the set of ideal answers.

Official metrics are based on the extended cumulated gain (XCG) (Kazai et al., 2004; Kazai and Lalmas, 2005). The XCG metrics are a family of metrics that aim to consider the dependency of XML elements (e.g. overlap and near misses) within the evaluation. The XCG metrics include the user-oriented measures of normalised extended cumulated gain (nXCG) and the system-oriented effort-precision/gain-recall measures (ep/gr). See (Kazai and Lalmas, 2005) for more details.

In the rest of the paper, runs will be evaluated using nXCG[10], nXCG[25], nXCG[50] and ep/gr MAP metrics, at it is the case for INEX official submissions. For a given rank i , the value of nXCG[i] reflects the relative gain the user accumulated up to that rank, compared to the gain he/she could have attained if the system would have produced the optimum best ranking.

4. Term weighting scheme

The choice of the term weighting scheme is not a trivial issue, since the term weighting scheme used should model the importance of terms in leaf nodes, but also in documents and in the whole collection. Term occurrences may follow other rules in XML documents than in flat text documents and weighting schemes traditionally used in IR should be adapted for XML retrieval. Many approaches use for example *ief* (*Inverse Element Frequency*), which models the importance of terms in the collection of elements. One can find in (Trotman, 2005) some examples of adaptation of traditional weighting schemes to XML retrieval.

Conclusions published in the literature are related to what is called in the INEX 2005 campaign *Thorough* strategy, i.e. a strategy which aims at finding all relevant elements. Experiments showed that the term weighting scheme should reflect the local importance of terms in elements but also their global importance in the collection (Sauvagnat et al., 2005; Trotman, 2005). We aim here at validate these conclusions for a *Focussed* strategy (the component length parameter is consequently introduced in other experiments).

4.1. Runs

We propose to evaluate the following weighting functions, which are used in equation [1]:

- We first test a very simple function, only based on term frequency in leaf nodes:

$$w_i^q = tf_i^q \quad w_i^{ln} = tf_i^{ln} \quad [5]$$

where tf_i^q and tf_i^{ln} are respectively the frequency of term i in query q and leaf node ln .

- In order to confirm the need to adapt the weighting function to the new granularity of information, we also test the well-known *tf.idf* function:

$$w_i^q = tf_i^q \times idf_i \quad w_i^{ln} = tf_i^{ln} \times idf_i \quad [6]$$

where $idf_i = \log(|D|/(|d_i|+1))+1$, with $|D|$ the total number of documents in the collection and $|d_i|$ the number of documents containing i .

- These functions are then adapted to take into account the new granularity of information we processed (we consider leaf nodes instead of whole documents):

$$w_i^q = tf_i^q \times ief_i \quad w_i^{ln} = tf_i^{ln} \times ief_i \quad [7]$$

where ief_i is the inverse element frequency: $ief_i = \log(|L|/(|ln_i|+1)) + 1$, with $|L|$ the total number of leaf nodes in the collection and $|ln_i|$ the number of leaf nodes containing i .

- These parameters are then combined to take into account of the importance of terms in both the collection of documents and the collection of leaf nodes.

$$w_i^q = tf_i^q \quad w_i^{ln} = tf_i^{ln} \times ief_i \times idf_i \quad [8]$$

Moreover, to determine the effect on relevance of the distance parameter in the propagation function (that allows to tune the length of preferred elements)- see equation [2]-, we experiment with values ranging from 0.1 (distance has a lot of importance) to 1 (distance has no importance) for the α parameter.

4.2 Results

Figure 1 and 2 show the evolution of the nXCG[10] and ep/gr evaluation metrics, for the generalised quantisation function.

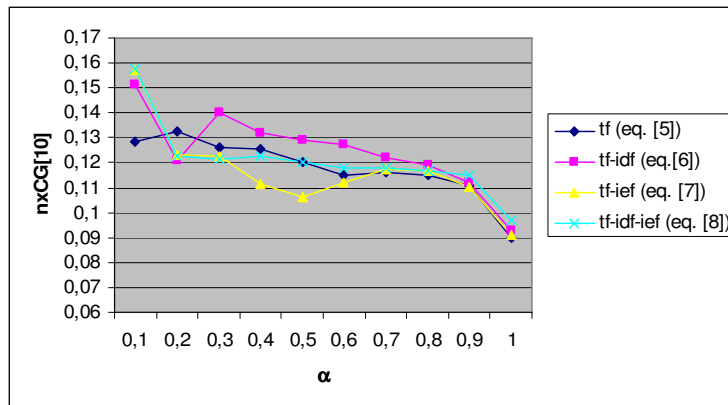


Figure 1: Evolution of nXCG[10] against α , generalized quantization function

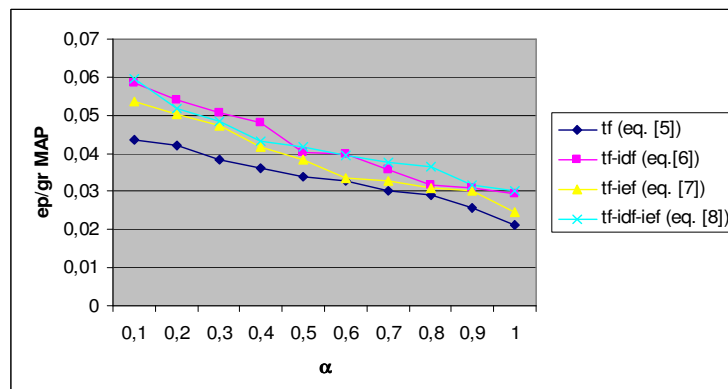


Figure 2: Evolution of ep/gr against α , generalized quantization function

Similar evolutions can be observed for nXCG[25] and nXCG[50] metrics and for the strict quantisation function.

Let us first comment results obtained with the different term weighting schemes. The simple use of the tf factor seems not to be enough for weighting terms. Using idf and ief factors significantly improves the results. Best results are obtained when combining the 3 factors (eq. [8]). This confirms previous results (Sauvagnat et al., 2005), in which we have shown that the weighting function used should take into account of the importance of terms in leaf nodes (tf factor), but also in the collection of elements (ief factor) and in the collection of documents (idf factor).

If we now consider the α parameter, we see that best results are obtained for small values of α , i.e. when small elements are preferred. This contradicts previous experiments on 2003 and 2004 topic sets (Sauvagnat and Boughanem, 2004), in which the optimal value of α was around 0.7. This can be explained by the new process used when doing the INEX 2005 assessments. Whereas in previous years assessors judged at the same time the exhaustivity and specificity dimensions of the returned elements, only the exhaustivity judgment was required in 2005. Assessors should select exhaustive parts of documents, and specificity was then inferred by the system collecting assessments (Lalmas and Piwowarski, 2005). This way, very specific elements are preferred by assessors, that explains our results.

In the rest of the paper, we will call *baseline run* the run performed using equation [8] ($tf-idf-ief$) as weighting formula and $\alpha = 0.1$ in the propagation function (equation [2]).

5. Introduction of component length

Results presented in the last section showed that when using small values of the α parameter, performances increase. Small values of the α parameter imply that very specific elements are preferred by the system. However, this may lead to return very small elements, i.e. elements that are not informative. Component length (i.e. the number of terms in a component) can be a crucial parameter to evaluate a component informativity. The problem is to know how and when introducing this parameter.

5.1 Experiments

In previous work, we have showed that the introduction of length in the term weighting scheme does not lead to better results in our model (Sauvagnat et al., 2005). In this paper, we propose to use component length (i) as a filter when ranking elements and (ii) during propagation to increase the importance of some particular nodes:

- (i) Element length can be used as a **threshold**, before returning elements.

To avoid the retrieval of nodes that do not supply information (like *title* nodes for example), we introduce the following rule:

Let n be a node and ln_i , $i \in [1..N]$ be its descendant leaf nodes having a non-zero relevance score. Let L be the sum of the length of ln_i (i.e. the sum of the number of terms contained in ln_i). If L is smaller than a given value x , n will be considered as not relevant.

This rule can be formalized as follows:

$$r_n = \begin{cases} |L_n| \sum_{k=1..N} \alpha^{dist(n, ln_k)-1} \times RSV(q, ln_k) & \text{if } L > x \\ 0 & \text{otherwise} \end{cases} \quad [9]$$

$$\text{Where } L = \sum_{i=1..N} l_i \text{ with } RSV(q, l_n) > 0 \quad [10]$$

and l_i is the length of leaf node l_n .

Smallest elements are consequently removed from the results list.

(ii) Node length can also be introduced by doing a so-called **weighted propagation**.

Intuitively, one can think that document writers use small node to highlight significant information. Those nodes can thus give precious indications on their ancestors' relevance. A *title* node in a *section* allows the subject of the *section* node to be better situated. We propose thus to emphasize small nodes role during the propagation.

Let l_k be the length of leaf node l_n and Δ_l be the average leaf nodes length in the collection. If a leaf node l_n is small (i.e. smaller than the average length), the relevance r_{par} of its direct parent node should be reduced (the parent node only contain textual information arised from this leaf node). It should however have a more prominent role than other leaf nodes when evaluating the relevance of its ancestors nodes.

To summarize, we introduce the $\beta(l_n)$ parameter in nodes relevance evaluation:

$$r_n = \left| L_n^r \right| \sum_{k \in L_n} \alpha^{\text{dist}(n, l_k)-1} \times \beta(l_n) \times RSV(q, l_k) \quad [11]$$

with¹

$$\beta(l_n) = \begin{cases} l_k / \Delta_l & \text{if } \text{dist}(n, l_k) = 1 \text{ and } l_k < \Delta_l \\ \log(\Delta_l / l_k) & \text{if } \text{dist}(n, l_k) > 1 \text{ and } l_k < \Delta_l \\ 1 & \text{otherwise} \end{cases} \quad [12]$$

5.2 Results

Table 1 shows results obtained when using element length as a threshold.

		nXCG[10]	nXCG[25]	nXCG[50]	ep/gr - MAP
Generalized	Baseline	0.158	0.1615	0.1592	0.0596
	x=10	0.1651	0.1733	0.1635	0.0559
	x=20	0.1602	0.1740	0.1605	0.0526
Strict	Baseline	0.0615	0.1002	0.1067	0.0243
	x=10	0.0841	0.1273	0.1463	0.0362
	x=20	0.0742	0.1145	0.1378	0.0345

Table 1: Results using a length threshold

We see that when using a small threshold (x=10), performances increase significantly (specially with the strict quantization function). It refutes conclusions drawn in (Kamps et al., 2004) and (Sauvagnat et al., 2005). This is probably due to the new methods used in 2005 for the assessments: some elements could be judged as too small, which was not the case for the other evaluation campaigns.

¹ This optimal β values has been obtained in an experimental way

Figure 3 shows the evolution of performance when doing a weighted propagation.

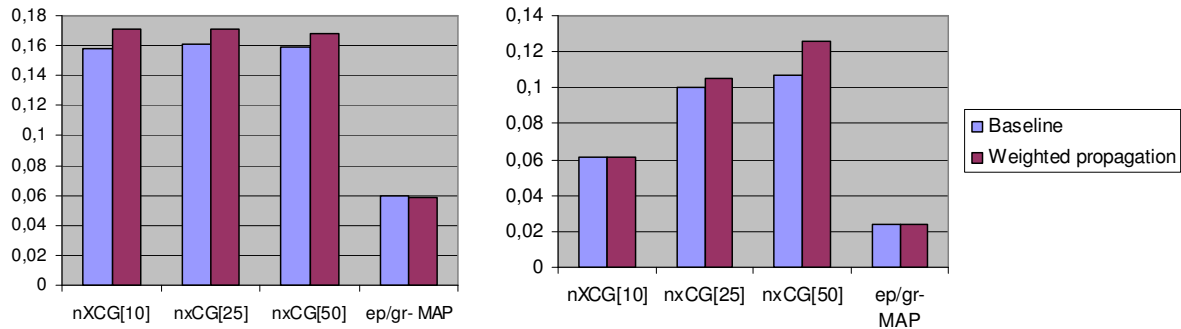


Figure 3: Evolution of all metrics with weighted propagation, generalized quantization (left) and strict quantization (right)

We see the interest of using information carried out by small elements during the propagation (up to 17% of performance increase): very small elements should not be returned to user, but can be useful for evaluating their ancestors' relevance.

At last, we should also notice that combining a length threshold and weighted propagation also improve results up to 20% and that performances are slightly better than those obtained when taking the two methods separately.

6. Contextual relevance

Former experiments (Sauvagnat et al., 2006b; Sigurbjörnsson et al, 2003; Mass and Mandelbrod, 2004; Arvola and al., 2005) have shown the interest of using contextual relevance in the evaluation of elements relevance. Documents can often be seen as entities, since authors create them by following some unity of thought. Context can consequently gives hints about document components relevance. Context can be interpreted in different ways: it can concern ancestors of elements or whole documents. In this study, we focus on context as document relevance.

We think that improvements induced by contextual relevance can be twofold: first, contextual relevance can help to find the appropriate granularity of elements (and can consequently play a role on the specificity dimension) and second, it can help to better identify document components' topics (and interfere on the exhaustivity dimension).

6.1 Experiments

Contextual relevance as document relevance can be introduced in two different ways in our model:

- 1- A first way is to *combine relevance of elements and relevance of documents*, by using document relevance when evaluating inner nodes relevance scores.

For this purpose, leaf nodes weights are propagated upwards in the document tree by using equation [2], until the root is assigned a relevance value r_{root} . We then use the following

propagation function for evaluating inner nodes relevance values, inspired from work presented in (Mass and Mandelbrod, 2004)² :

$$r_n = \rho \cdot |L_n^r| \cdot \sum_{\ln_k \in L_n} \alpha^{dist(n, \ln_k)-1} \times RSV(q, \ln_k) + (1 - \rho) \times r_{root} \quad [13]$$

with r_{root} the relevance of the root node of the document, evaluated with equation [2]. $\rho \in [0..1]$ is a parameter used as pivot, that allows to fit the importance of root node relevance in inner nodes relevance evaluation. The use of ρ can be seen as doing a *backwards propagation* of document relevance in the document tree.

- 2- To evaluate the relevance of elements and documents separately, we propose to rank elements using these two criteria: elements score and documents score. More precisely, elements are first **ranked by the relevance of the document** they belong to, and then by their own relevance.

We use the following algorithm:

- relevance values are computed for each document in the collection, using the Mercure system (Boughanem et al., 1998);
- relevance values are computed for each node of the collection, using equation [2]
- documents are ranked by decreasing order of relevance;
- for each document, elements they contain are ranked by decreasing order of relevance and are returned to users.

6.2. Results

Backwards propagation.

We experiment with values of ρ ranging from 0.1 to 1 (no backwards propagation). Figure 4 shows the evolution of results with the nXCG metric.

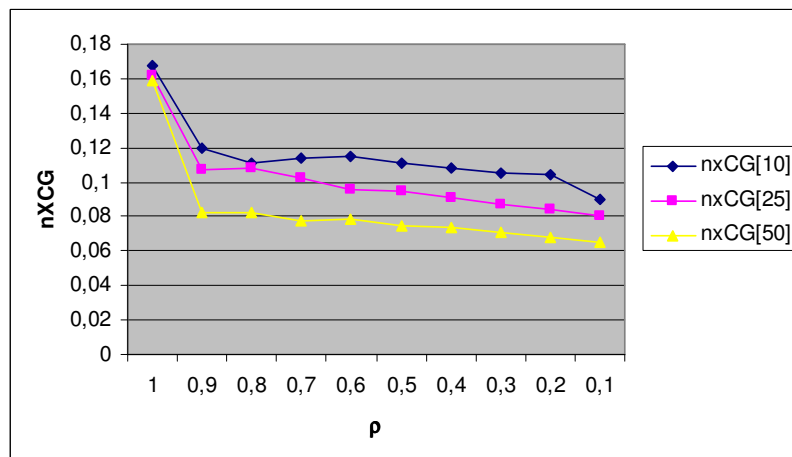


Figure 4: Evolution of nXCG metrics against ρ , generalised quantisation function

² Our formula differs from the ones in (Sigurbjörnsson et al, 2003; Mass and Mandelbrod, 2004) since relevance values of root and inner nodes are computed in a different way

Ranking on document relevance.

Figure 5 shows the evolution of results with all metrics when ranking on document relevance.

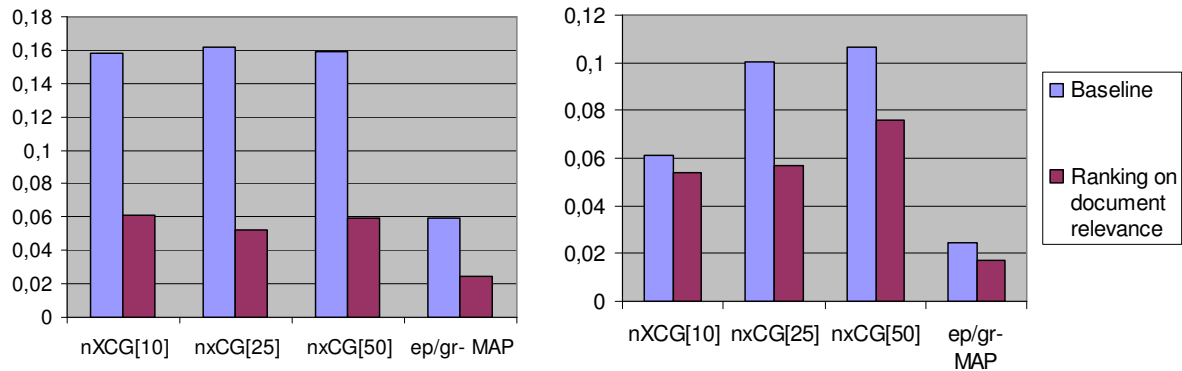


Figure 4: Evolution of all metrics when ranking on document relevance, generalized quantization (left) and strict quantization (right)

Comments.

As opposed to results obtained (Sauvagnat et al., 2006b; Sigurbjörnsson et al, 2003; Mass and Mandelbrod, 2004; Arvola and al., 2005), the introduction of contextual relevance (with backwards propagation or by ranking elements on document relevance) has a negative effect on outcomes. This result can be explained by the fact that results elements should focus on the user information need and should be auto-explanatory. As they should also be considered independently, context information has no impact on relevance (which was not the case in a thorough strategy).

7. Introduction of structural hints

CO queries simulate a user who does not know (or does not want to use) the actual structure of the XML documents in a query.

However, users, when expressing CO queries, may decide to add structural hints in their queries, in order to avoid systems returning too many irrelevant hits. Such hints should not be satisfied strictly, but should be considered as clues on which types of elements are most probably relevant. Such structural hints were added in INEX 2005 CO queries. These "new" CO queries were called CO+S queries.

During the assessments, structural conditions were ignored. Judges assessed the elements returned for CO+S queries as whether they satisfied the information need with respect to the content criterion only.

We propose here to test if structural hints improve overall performance of our system.

6.1. Experiments

Nodes relevance are evaluated as follows. INEX CO+S queries are of the form: $tg[q]$ where tg is a tag name, i.e. a structure constraint, and $q=t_1, \dots, t_n$ is a content constraint (i.e. a CO query) composed of simple keywords terms. The relevance value of a node n to a CO+S query is computed according to the following formula:

$$r_n = \begin{cases} \sum_{ln_k \in L_n} \alpha^{dist(n, ln_k)-1} \times RSV(q, ln_k) & \text{if } n \in \text{construct}(tg) \\ 0 & \text{otherwise} \end{cases} \quad [14]$$

where the $\text{construct}(tg)$ function allows the creation of a set composed of nodes having tg as tag name, and $RSV(q, ln_k)$ is evaluated with equation [1].

The $\text{construct}(tg)$ function uses a Dictionary index, which provides for a given tag tg the tags that are considered as equivalent. This index is built manually. For processing CO+S queries, as structural conditions should be considered as vague conditions, we use a dictionary index composed of very extended equivalencies. For example, a section node (sec) can be considered as equivalent to both a paragraph (p) and a body (bdy) node.

6.2 Results

Table 2 shows the results obtained when using structural hints in CO queries.

		nXCG[10]	nXCG[25]	nXCG[50]	ep/gr - MAP
Generalized	Baseline	0.158	0.1615	0.1592	0.0596
	CO+S	0.237	0.198	0.1927	0.0752
	Gain	+50%	+22%	+21%	+26%
Strict	Baseline	0.0615	0.1002	0.1067	0.0243
	CO+S	0.0841	0.1273	0.1463	0.362
	Gain	+36%	+27%	+37%	+48%

Table 1: Results using a length threshold

Results are significantly better (up to 50% increase) when structural hints are used, for all levels of recall and when considering mean average precision. This confirms results in (Sauvagnat et al., 2006) and is not surprising, since the user need is more clearly defined. This however contradicts results obtained in (Kamps et al., 2005), in which authors showed that structured queries do not lead to improved mean average precision, but that structured queries function as a precision enhancing device: they are useful for promoting the precision in initially retrieved elements, but are not useful on mean average precision.

8. Discussion and Conclusion

We can identify a number of important points to the experiments presented above. For a focussed retrieval strategy, a term weighting scheme taking into account the importance of terms in elements and both in collection of elements and collection of documents is useful. Moreover, the introduction of component length as a threshold on results or used in a weighted propagation function improves significantly the results. Third, contextual relevance seems not to be useful which contradicts results obtained by state-of-the-art methods for non-focussed

retrieval. At last, the use of structural hints increases up to 50% results we obtained when using queries composed only of simple keyword terms.

We have to notice the relative high precision of our runs comparing to INEX 2005 official metrics (Fuhr et al., 2005). Most of our runs would have been ranked in the top ten for strict quantization when using component length. Best results are however obtained when using structural hints (which corresponds to the COS.Focussed retrieval task at INEX 2005). We would have been ranked first for generalised quantisation on nXCG[10], nXCG[25] and nXCG[50] metrics. We respectively obtain 0.237, 0.198 and 0.1927, whereas best results were respectively 0.2181, 0.1918 and 0.1817, and were obtained by the University of Amsterdam with a language model-based method (Sigurbjörnsson et al, 2005) and by IBM Haifa Research Lab (Mass and Mandelbrod, 2005) using the vector space model. We would also be in the top 5 for strict quantization.

Let us say a few words about model dependency. The retrieval strategies we proposed are only tested with our retrieval system. However, as we obtain similar conclusions as state-of-the-art methods in thorough retrieval, we expect that our conclusions on focussed retrieval can be generalized to other focussed retrieval methods.

Moreover, our conclusions are of course collection-dependent, assessments-dependent and metrics-dependent. We know that the XCG metrics are not perfect (Geva, 2005) since they do not penalize results having overlap elements, but they have the main (and very important) advantage to allow the evaluation of focussed strategies. Moreover, the relevance assessments procedure had changed in 2005, that probably advantages specificity against exhaustivity, and consequently advantages the retrieval of smaller nodes than the other years. This is thus hard to compare results with results obtained past years. This is mainly true on conclusions about the use of element length. Since user tasks seem now to be clearly defined, we hope having consistent conclusions in the future, which will be confirmed by significance tests. At last, experiments are currently running on the INEX 2006 collection (which is composed of Wikipedia articles) to generalize our results.

9. References

- Abolhassani, M., and Fuhr, N., (2004). Applying the divergence from randomness approach for content-only search in XML documents. In *Proceedings of ECIR 2004*, Sunderland, pages 409–419.
- Anh, V. N. and Moffat, A.(2002). Compression and an IR approach to XML retrieval. In *Proceedings of INEX 2002 Workshop*, Dagstuhl, Germany.
- Arvola, P. , Junkkari, M. , and Kekalainen, J.(2005). Generalized contextualization method for XML information retrieval. In *Proceedings of CIKM 2005*, Bremen, Germany.
- Boughanem, M., Dkaki, T. , Mothe, J. , Soule-Dupuy, C. (1998). Mercure at TREC-7. In *Proceedings of TREC-7*, 1998.
- Fuhr, N., Grossjohann, K. (2004) XIRQL: an XML query language based on information retrieval concepts, *ACM Transactions on Information Systems*,22, pages 313-356.
- Fuhr, N. , Lalmas, M. , Malik, S. , Kazai, G. (2005). *INEX 2005 workshop proceedings*.
- Geva, S.. XCG overlap at INEX 2004. In *Proceedings of INEX 2005*, Dagstuhl, Germany.
- Gövert, N. , Abolhassani, M. , Fuhr, N. , and Grossjohann, K. (2002). Content-oriented XML retrieval with HYREX. In *Proceedings INEX 2002*, Dagstuhl, Germany.
- Grabs, T. and Scheck, H.-J.(2002). Flexible information retrieval from XML with PowerDB XML. In *Proceedings of INEX 2002, Dagstuhl, Germany*, pages 26–32.
- Kakade, V. and Raghavan, P. (2005). Encoding XML in vector spaces. In *Proceedings of ECIR 2005*, Saint Jacques de Compostelle, Spain, 2005.

- Kamps, J. , de Rijke, M., Sigurbjörnsson, B. (2004) Length normalization in XML retrieval. In *Proceedings of SIGIR 2004*, Sheffield, England, pages 80–87.
- Kamps, J. , Marx, M., de Rijke, M. , and Sigurbjörnsson, B.(2005). Structured queries in XML retrieval. In *Proceedings of CIKM 2005*, Bremen, Germany, 2005.
- Kazai, G., Lalmas, M., and de Vries, A. P.(2004). The overlap problem in content-oriented XML retrieval evaluation. In *Proceedings of SIGIR 2004*, Sheffield, England, pages 72–79.
- Kazai, G. and Lalmas, M. (2005). INEX 2005 evaluation metrics. In *Proceedings of INEX 2005*, Dagstuhl, Germany.
- Kekalainen, J. , Junkkari, M. , Arvola, P. , and Aalto T.(2004). Trix 2004 struggling with the overlap. In *INEX 2004 Proceedings*, Dagstuhl, Germany.
- Lalmas, M. and Piwowarski, B.(2005) . INEX 2005 relevance assessment guide. In *INEX 2005 Proceedings*, pages 391–401.
- Mass, Y. and Mandelbrod, M.(2004) Component ranking and automatic query refinement for XML retrieval. In *Proceedings of INEX 2004*, pages 134–140.
- Mass, Y. and Mandelbrod, M.(2005). Experimenting various user models for XML retrieval. In *Proceedings of INEX 2005*, Dagstuhl, Germany.
- Sauvagnat, K. and Boughanem, M. (2004). Using a relevance propagation method for Adhoc and heterogeneous tracks at INEX 2004, In *Proceedings of INEX 2004*, Dagstuhl, Germany.
- Sauvagnat, K., Hlaoua, L., and Boughanem, M. (2005). XFIRM at INEX 2005: adhoc and relevance feedback tracks. In *Proceedings of INEX 2005* , Dagstuhl, Germany.
- Sauvagnat, K. and Boughanem, M. and Chrisment, C., (2006). Why using structural hints in XML retrieval ?, In *Proceedings of Flexible Query Answering (FQAS) 2006* , Milano, Italia.
- Sauvagnat, K., Hlaoua, L., Boughanem, M. (2006b), XML retrieval: what about using contextual relevance? In *proceedings of ACM Symposium on Applied Computing (SAC) - IAR (Information Access and Retrieval)* , Dijon, France.
- Sigurbjörnsson, B. , Kamps, J. , de Rijke, M.. An element-based approach to XML retrieval. In *Proceedings of INEX 2003 workshop*, Dagstuhl, Germany.
- Sigurbjörnsson, B. , Kamps, J. , de Rijke, M. (2005). The university of Amsterdam at INEX 2005: Adhoc track. In *INEX 2005 Proceedings*, Dagstuhl, Germany.
- Trotman, A.(2005). Choosing document structure weights. *Information Processing and Management*, 41(2):pages 243–264.
- Zwolf, R. van , Wiering, F. , and Dignum, V.(2004). The Utrecht blend: basic ingredients for an xml retrieval system. In *INEX 2004 proceedings*, Dagstuhl, Germany.