

XML retrieval: what about using contextual relevance?

Karen Sauvagnat
IRIT/SIG
118 route de Narbonne
F-31 062 Toulouse Cedex 4,
France
sauvagna@irit.fr

Lobna Hlaoua
IRIT/SIG
118 route de Narbonne
F-31 062 Toulouse Cedex 4,
France
hlaoua@irit.fr

Mohand Boughanem
IRIT/SIG
118 route de Narbonne
F-31 062 Toulouse Cedex 4,
France
boughane@irit.fr

ABSTRACT

The aim of this study is to evaluate the impact of context to better identify relevant elements in XML retrieval. Context is represented here by clues on whole document relevance. We represent context according to different points of view: by introducing document dimension while computing terms weights, by using document relevance when evaluating elements relevance or by ranking elements on document relevance. Experiments were undertaken on INEX collection, and results showed the interest of contextual relevance and a relative high precision of our proposal comparing to INEX official results.

Keywords

XML retrieval, relevance propagation, contextual relevance

1. INTRODUCTION

XML IRS aim at finding relevant *document components* (also called *information units*, *nodes*, *elements* or *subtrees*) instead of relevant documents. The aim is not to find an entry point in the document, but on the contrary to return an information unit that focusses on the user information need and that does not depend on another to be understood.

We propose here to consider contextual relevance to help finding appropriate granularity of document components and to better understand their content. The main idea is: an element in a relevant context should be ranked higher than an identical element in a non-relevant context. In XML documents, context of a given element is provided by its ancestors. Context can consequently be viewed with different levels of granularity. For this first study, we restrict the notion of context to whole document. This restriction is motivated by the fact that whole documents are often considered as entities for both authors and users. *Our aim is to revisit the context idea by introducing it in 3 different ways, and to provide a comparative analysis of such methods.*

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. To copy otherwise, to republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee.

SAC'06 April 23-27, 2006, Dijon, France

Copyright 2006 ACM 1-59593-108-2/06/0004 ...\$5.00.

2. EXPERIMENTAL SETUP

In order to evaluate the introduction of context in element relevance evaluation, we compare results to those obtained with a background model based on a relevance propagation method [3]. In this model, queries are processed as follows: relevance values are assigned to leaf nodes and relevance score of inner nodes are then computed dynamically, thanks to a propagation of leaf nodes score through the document tree. An ordered list of subtrees is then returned to the user.

We evaluate the introduction of contextual relevance in this background model in three different ways.

- 1- First, contextual relevance is introduced at *leaf nodes level*, when taking into account *term weights*.

Indeed, evaluation of term weights can reflect the importance of terms in leaf nodes, but also in whole documents. We consequently evaluate various term weighting schemes, by redefining or not the *idf* factor. Weights w_i^{ln} and w_i^q of term i in leaf node ln and query q can either be defined as:

$$w_i^q = tf_i^q * idf_i \quad w_i^{ln} = tf_i^{ln} * idf_i \quad (1)$$

$$w_i^q = tf_i^q * ief_i \quad w_i^{ln} = tf_i^{ln} * ief_i \quad (2)$$

$$w_i^q = tf_i^q \quad w_i^{ln} = tf_i^{ln} * idf_i * ief_i \quad (3)$$

Where tf_i^{ln} and tf_i^q are respectively the frequency of i in ln and q , idf_i is the inverse document frequency of i , ief_i is the inverse element frequency of i , i.e. $\log(|N|/|nf_i| + 1) + 1$, where $|nf_i|$ is the number of leaf nodes containing i and $|N|$ is the total number of leaf nodes in the collection.

- 2- A second way to introduce contextual relevance is to *combine relevance of elements and relevance of documents*, by using document relevance when evaluating inner nodes relevance scores.

For this purpose, we propagate leaf nodes weights upwards in the document tree until the root is assigned a relevance value r_{root} . We then use the following function for evaluating inner nodes relevance values, inspired from work presented in [2]. The relevance r'_n of node n is expressed as:

$$r'_n = \rho * r_n + (1 - \rho) * r_{root} \quad (4)$$

with r_{root} the relevance of the root node of the document and r_n the first relevance of node n . Both are evaluated thanks to the propagation function presented in [3]. $\rho \in [0..1]$ 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.

- 3- 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.

Collection, topics and metrics. Evaluation is done with INEX 2003 and INEX 2004 CO (Content-Only) topics sets. The INEX metrics for evaluation are based on the traditional recall and precision measures. To obtain recall/precision figures, the two dimensions of relevance (exhaustivity and specificity) need to be quantised into a single relevance value. Quantisation functions for many user standpoints are available. We used here an average of all quantisation functions, called *MAP* (*Mean Average Precision*) for evaluating general performance.

3. RESULTS AND DISCUSSION

Term weighting scheme. Table 1 shows results obtained when using various term weighting schemes. We can notice that introducing term importance in both collection of leaf nodes and collection of documents ($tf*idf*ief$ formula) improves results on both topic sets, and in a very significant way for 2004. Introduction of *idf* seems in this last case to be prominent for explaining such results.

Table 1: Comparison of term weighting schemes

	2003	2004
$tf-idf$ (eq. 1)	0.1204	0.1045
$tf-ief$ (eq. 2)	0.1219	0.0910
$tf-idf-ief$ (eq. 3)	0.1230	0.1084

Combining relevance of documents and relevance of elements. In experiments on backwards propagation (equation 4), we use equation 2 for computing term weights. Indeed, the aim here is to evaluate the impact of introduction of contextual relevance at inner nodes evaluation level, and document relevance should consequently not be introduced before. Figure 1 shows the general evolution of precision against ρ . We observe a clear increase of average precision when contextual relevance is taken into account for the evaluation of nodes relevance values (up to 30% for the 2004 topic set). Nodes context should however not be too important in the propagation function, since optimal values for ρ are relatively high.

Ranking on document relevance. Experiments in this section first rank nodes according to the relevance of their associated document, and then by their own relevance value. We evaluate the document relevance in two different ways: (i) by using a simple $tf-idf$ formula, and (ii) using the propagation function defined in [3].

Results in table 2 are compared to results obtained by doing a simple propagation and by ranking nodes according to their relevance value. While performances decrease on 2003 topic set when doing a first ranking on document relevance, an opposite outcome is observed for the 2004 topic set. In fact, we notice up to 40% of precision increase when a first ranking is done on document relevance.

4. DISCUSSION AND CONCLUSIONS

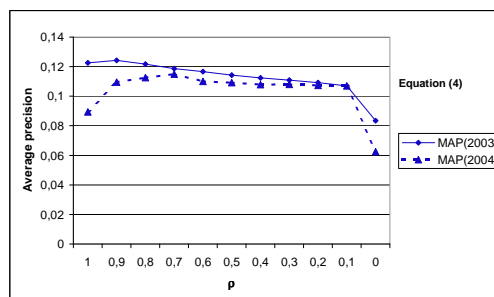


Figure 1: Evolution of average precision against ρ

Table 2: Comparison of average precision obtained when ranking elements on document relevance

	2003	2004
<i>Baseline</i>	0.1219	0.0910
<i>tf-idf on document</i>	0.1066	0.1204
<i>Propagation on document</i>	0.1033	0.1073

Whatever the contextual relevance we used, we showed its interest on exhaustivity and specificity. Even if a document contains heterogeneous parts (as it's the case in the INEX collection), a semantic unity can be found.

Introducing the *idf* factor in the term weighting scheme (equation 3), as well as using document relevance for computing element relevance (equation 4) increases overall performance. Experiments combining both approaches (not presented here due to space limitation) show an increase of average precision compared to approaches used separately, especially on 2004 topic set.

Ranking on document relevance also allows the increase of overall effectiveness, especially on 2004 topic set. Average precisions obtained can however be compared to those obtained using backwards propagation (eq. 4). As backwards propagation can also improve results on 2003 topic sets, it seems to be a better way to introduce contextual relevance. Some results are contradictory between 2003 and 2004 topic sets, which seems to highlight a problem in relevance assessments. In fact, according statistics edited in [1] only 12% of non-zero agreement is obtained on queries owning duplicate assessments (!). Assessments should be more clearly defined in order to have consistent results.

At last, we have to notice the relatively high precision of our runs comparing to official INEX results. We would have been ranked in the top 5 for almost all quantisation functions. We now plan to evaluate the impact of ancestors' relevance on overall performance. Since documents relevance is useful for improving results, what about the relevance of different ancestors at different levels of the tree hierarchy?

5. REFERENCES

- [1] M. Lalmas and al. Some statistics about INEX 2004. INEX 2004 Workshop, 2004.
- [2] Y. Mass and M. Mandelbrod. Component ranking and automatic query refinement for XML retrieval. In *Proceedings of INEX 2004*, Springer, 2005.
- [3] K. Sauvagnat and M. Boughanen. Using a relevance propagation method for adhoc and heterogeneous tracks. In *Proceedings of INEX 2004*, Springer, 2005.