Tuning and Evolving Retrieval Engine by Training on Previous INEX Testbeds

Gilles Hubert

IRIT/SIG-EVI, 118 route de Narbonne, F-31062 Toulouse cedex 9
hubert@irit.fr

Abstract. This paper describes the retrieval approach proposed by the SIG/EVI group of the IRIT research centre at INEX'2006. This XML approach is based on direct contribution of the components constituting an information need. This paper focuses on the impact of changes between INEX'2005 and INEX'2006 notably the corpus change. This paper describes the search engine configurations and evolutions resulting from training on previous INEX testbeds and used to participate to INEX'2006. It presents also the results of the official experiments carried out at INEX'2006 and additional results.

1 Introduction

Since the beginning of the INEX initiative, various XML retrieval systems and various evolutions of these systems were proposed. XML retrieval needs to take into account both content and structural aspects. XML retrieval systems can be separated into information retrieval systems based on probabilistic models [8][11][12][15], information retrieval systems based on vector space models [2][4][6][9] and systems based on databases [3][10].

A framework such as INEX is useful to try to estimate a global effectiveness of a system and to determine the contexts adapted to a system. Among the systems that participated to INEX previous years and that obtained globally good results there are approaches based on vector space models such as [4][9], probabilistic methods such as [11][15] and database systems such as [10] depending on task and quantisation.
In this paper, we present an IR method based on a vector space model. However, this approach is based on direct contribution of each component of the query and particularly on the presence of each term constituting the query. The method meets other proposals such as [4][13] in some principles but differs from heuristics, score aggregation principle or XML structure management. Different parameters are intended to provide configuration possibilities to adapt the method notably according to task and quantisation.

In the remainder of this paper, Section 2 summarizes the objectives of this year participation to the INEX'2006 round. Then, a presentation of the guiding principles on which relies the retrieval method is done in Section 3. Section 4 details the submitted runs and the official and additional results. Section 5 concludes this paper.

N. Fuhr, M. Lalmas, and A. Trotman (Eds.): INEX 2006, LNCS 4518, pp. 243–252, 2007.

2 Participation Objectives

Participating to INEX this year had multiple objectives:

- On the one hand the interest was to estimate the influence of changes introduced in the INEX 2006 framework regarding corpus and tasks. The new Wikipedia corpus has features different from the past IEEE corpus notably regarding corpus size, document contents and document structures. Furthermore, a new task has been defined i.e. Best in Context. In addition, runs using queries on content only and runs using queries on content and structure were merged for evaluation. This new processing offers a mean to evaluate whether treating structural hints in our method improves results or not,
- On the other hand, the interest was to study the behaviour of different configurations of our method resulting from learning on previous INEX testbeds. These configurations intend to be suited to the different tasks and quantisations defined in INEX.

3 Method Principles

The IR method described in this paper is based on a vector space model. Document and query representations are comparable to vectors. However, the correspondence between documents and query is not estimated using a “usual” similarity measure. The method is based on a generic scoring function that can be adapted to different retrieval contexts. The current definition of the scoring function results from work on automatic document categorization [1] and work on XML retrieval at previous INEX rounds [6][7][14]. The scoring function is based on direct contribution of each query term appearing in an XML element. The contribution can be modulated according to other components of the query such as structural constraints. A principle of score aggregation completes the method with regard to the hierarchical structure of XML documents.

The scoring function is defined as a combination of three values. It can be globally defined as follows:

$$Score(Q,E) = \left(\sum_i f(t_i,E) \cdot g(t_i,Q) \right) \cdot h(Q,E)$$

Where

Q is the query

t_i is a term representing the query Q

E is an XML element

$f(t_i,E)$ This factor estimates the importance of the term t_i in the XML element E.

$g(t_i,E)$ This factor estimates the importance of the term t_i in the query representation Q.

$h(Q,E)$ This factor estimates the global presence of the query Q in the XML element E.

On the one hand, the function is defined as an addition of contributions of the concepts constituting a query. This principle allows giving relevance to elements dealing about either only one concept or several concepts. The addition tends to promote elements containing several concepts. However, depending on the different chosen functions an element dealing strongly about one concept can be estimated higher than an element dealing lightly about many concepts. On the other hand, the function estimates globally the relevance of an element according to a query.

The function *f* that estimates the importance of a term in an XML element is based on the number of occurrences of the term in the element moderated by the number of XML elements of the corpus containing the term. Using this latter factor, the function increases the contributions of terms appearing in few XML elements of the corpus. This principle is similar to the tf.idf principle. A coefficient related to structural constraints on content term intends to increase or reduce term contributions according to constraint matching.

The function *g* that estimates the importance of a term in the query representation is based on the frequency of the term in the topic. The frequency is moderated by the total number of occurrences of terms in the query. A coefficient related to term prefixes intends to increase or reduce term contributions according to sign '+' and '-' associated to terms in the query.

The function *h* that estimates the global presence of a query in an XML element is based on the proportion of terms common to the query and the element with respect to the number of distinct query terms. A function power is used to clearly distinguish the elements containing a lot of terms describing the query from the elements containing few terms of the query.

So, the scoring function is defined as follows:

$$Score_t(Q,E) = \left(\sum_i \frac{cc(t_i,E) \cdot Occ(t_i,E)}{NbE(t_i)} \cdot \frac{prf(t_i,Q) \cdot Occ(t_i,Q)}{Occ(Q)} \right) \cdot \varphi^{\left(\frac{NbT(Q,E)}{NbT(Q)}\right)}$$

where

t_i is a term representing the query Q

E is an XML element

$cc(t_i,E)$ Coefficient defined for the matching of constraint on content (associated to the term t_i) by the element E.

$Occ(t_i,E)$ Number of occurrences of the term t_i in the element E.

$NbE(t_i)$ Number of elements containing the term t_i

$prf(t_i,Q)$ Coefficient defined for the prefix associated to the term t_i in the query Q.

$Occ(t_i,Q)$ Number of occurrences of the term t_i in the query Q.

$Occ(Q)$ Total of occurrences of all the terms representing Q.

φ Query presence coefficient, positive real

NbT(Q,E) Number of terms of the query Q and that appear in the XML element E.

NbT(Q) Number of distinct terms of the query Q.

The coefficients $cc(t_i,E)$ and $prf(t_i,Q)$ can be defined by functions. At the moment, these coefficients are defined as follows:

```
cc(t_i,E)    if E does not match the structural constraint defined on $t_i$
             then  $cc(t_i,E)=0.5$
             else  $cc(t_i,E)=1.0$

prf(t_i,Q)   if prefix is '+'
             then  $prf(t_i,Q)=5.0$
             else  if prefix is '-'
                   then $prf(t_i,Q)=-5.0$
                   else  $prf(t_i,Q)=1.0$
```

This solution allows attaching variable importance to structural constraints on content and prefixes. These definitions are resulting from experiments carried out on INEX'2003 and INEX'2004 testbeds.

The hierarchical structure of XML is taken into account through score aggregation. The hypothesis on which is based our method is that an element containing a component selected as relevant is also relevant and more if it has several relevant components. So, in our approach the score of an element is defined as the sum of its score computed according to its textual content (if it exists) and the scores of its descendant components that have a textual content (if they exist). The score of a component can be modulated (for example, according to the distance between the component and the ascendant) when aggregating in the ascendant depending on the applied strategy. At the moment, the aggregation is defined as follows:

$$Score_a(Q,E) = Score_t(Q,E) + \sum_l \alpha^{\frac{d(E,E_l)}{d(E_r,E_l)}} \cdot Score_t(Q,E_l)$$

where

α (real) is the score aggregation coefficient

E, E_l and E_r are XML elements

E_l is a descendant component of the E structural hierarchy (document) such as E_l has textual content

E_r is the root element of the structural hierarchy (document) of which E is a descendant component

d(X,Y) is the distance between an element X and its descendant element Y (for example in the path /article/bdy/sec/p[2], d(bdy, p[2]) = 2).

The coefficient α allows varying the influence of scores of descendant components in the aggregated score of an XML element. Leaf components have no descendant thus for such components: $Score_a(T,E) = Score_t(T,E)$.

Two types of structural constraints can be used to define INEX topics:

- constraints on content (e.g. about(.//p,'+XML +"information retrieval"),
- constraints on the granularity of target elements (e.g //article[….]).

As seen above, structural constraints on content are taken into account adding a coefficient cc(t_i,E) in the scoring function $Score_t$.

Structural constraints on the granularity of target elements are handled adding a coefficient that modifies the aggregated score $Score_a$ (equal to $Score_t$ for leaf nodes). The general principle is that if the XML element does not verify the constraint on target granularity associated to the query, the score computed is reduced. The aggregated score including granularity coefficient is therefore defined as follows:

$$Score_a(Q,E) = gc(Q,E) \cdot \left(Score_t(Q,E) + \sum_l \alpha^{\frac{d(E,E_l)}{d(E_r,E_l)}} \cdot Score_t(Q,E_l) \right)$$

where 3

$gc(Q,E)$ Coefficient defined for the matching of constraint on target (associated to the query Q) by the element E.

At the moment, this coefficient is defined as follows:

if E does not match the structural constraint defined on Q
then gc(Q,E)=0.5
else gc(Q,E)=1.0

This definition results from experiments carried out on INEX'2003 and INEX'2004 testbeds.

This solution allows attaching variable importance to structural constraints on result granularity. When gc(Q,E)=0.0 for elements that do not match the structural constraints, the structural constraints on result are strictly taken into account. When gc(Q,E)=1.0 the structural constraints on result are not taken into account.

The scoring function is completed by the notion of coverage. Coverage is a threshold corresponding to the percentage minimum of query terms that have to appear in an element to select it. It aims at ensure that only documents in which the query is represented enough will be selected for this topic. Coverage is combined to the scoring function as follows:

$$\text{If} \quad \frac{NbT(Q,E)}{NbT(Q)} < CT \quad \text{Then} \quad Score_a(Q,E) = 0.0$$

Where

CT	Coverage threshold, real positive, 0.0≤CT≤1.0
$NbT(Q,E)$	Number of terms common to the query Q and the element E
$NbT(Q)$	Number of distinct terms describing the query Q

Coverage is currently combined to Score$_a$. Otherwise, it can be applied to Score$_t$ and then it has consequences on aggregated scores.

An additional process can be done to eliminate overlapping elements in the result. This process consists in filtering the result and keeping according to a defined strategy only one element when two overlapping elements are encountered. A strategy is for example to keep the element with the highest score.

4 Experiments

At least two runs based on our XML retrieval method were submitted to INEX 2006 for each subtask one using the title part of queries and the other using the castitle part. Depending on the subtask, an additional run using either title or castitle was submitted.

4.1 Experiment Setup

Resulting from experiments and learning when participating to INEX'2003 and INEX'2004, all submitted runs shared the same values for the prefix coefficient prf(t_i,Q) and the coverage threshold CT (cf section 3). The coverage threshold was fixed to 0.35 (i.e. more than a third of terms describing the topic must appear in the text to keep the XML component). The values of prefix coefficient applied were fixed to +5.0 for the prefix '+', -5.0 for the prefix '-' and 1.0 for not prefixed terms.

For all the subtasks, CAS runs (i.e. having a label containing 'CAS') used the castitle part of each topic definition to define queries instead of the title part used for other runs. The coefficients taking into account structural constraints were fixed to 0.5 (i.e. the contribution of a query term is divided by 2 when the element does not meet the structural constraint) for all the subtasks. Structural constraints were handled so as vague conditions.

Furthermore, depending on the subtask we studied three combinations of score aggregation coefficient α and query presence coefficient φ (cf section 3) resulting from learning and experiments on INEX'2005 testbeds. The following combinations were tested:

- runs with labels containing ‘CH01x1’ used $\alpha=0.1$ and $\varphi=1$. This combination obtained good results on INEX’2005 testbeds for the subtask Thorough and the quantisation strict. This combination includes weak score aggregation and does not consider global query presence.
- runs with labels containing ‘CH06x50’ used $\alpha=0.6$ and $\varphi=50$. This combination obtained good results on INEX’2005 testbeds for the subtask Thorough and the quantisation generalised. This combination includes rather important score aggregation and considers moderately global query presence.
- runs with labels containing ‘CH0x3000’ used $\alpha=0.0$ and $\varphi=3000$. This combination obtained good results on INEX’2005 testbeds for the subtask Focused notably for strict quantisation. This combination does not include score aggregation and considers strongly global query presence.

4.2 Official and Additional Results

The official results corresponding to different configurations of our method tested for the different are detailed in the following tables. Additional results are included for the Thorough task in Table 1.

Table 1. Results for the subtask Thorough

Task	Thorough							
Metric: ep/gr	*Quantisation: generalised, overlap=off*							
	Official results				*Additional results*			
Run	V2006Cg35CH01x1tho		V2006CASCH01x1tho		*Cg35CH06x50*		*Cg35CH05x1000*	
	Maep	Rank	Maep	Rank	*Maep*	*Rank*	*Maep*	*Rank*
	0.0147	47/106	0.0161	40/106	*0.0232*	*(23)*	*0.0259*	*(20)*

A first observation is that official results are slightly over average. The configuration results from experiments on strict quantisation with INEX’2005 testbeds and gave weaker results for generalised quantisation. Without official results for strict quantisation final conclusions cannot be established. The INEX’2006 official results tend to show same behaviours of configurations for generalised quantisation. The additional results notably for the run labelled *Cg35CH06x50*, which gave good results on INEX’2005 testbed for the same task and quantisation, lead to better evaluations with INEX’2006 data.

Another observation is that structural conditions seem to improve the results since the run using castitle parts of queries obtain higher average precision than the run with same configuration using titles of queries.

A first observation is that results are average. The results have the same behaviour as with INEX’2005 testbeds for the same task and quantisation. Another observation is that treating only elements with textual content without including score aggregation leads to better results. To return leaf nodes seems to be better for the Focused task than to return intermediate nodes.

Table 2. Results for the subtask Focused

Task	Focused					
Metric: nxCG	*Quantisation: generalised*					
Run	V2006CH0x3000foc		V2006CASCH0x3000foc		V2006CH06x50foc	
overlap=on	precision	rank	precision	rank	precision	rank
nxCG@5	0.2899	43/85	0.2848	53/85	0.2630	61/85
nxCG@10	0.2472	42/85	0.2435	46/85	0.2198	62/85
nxCG@25	0.1905	43/85	0.1843	48/85	0.1759	52/85
nxCG@50	0.1558	36/85	0.1472	42/85	0.1471	43/85
overlap=off	precision	rank	precision	rank	precision	rank
nxCG@5	0.3151	41/85	0.3118	43/85	0.2666	63/85
nxCG@10	0.2841	35/85	0.2742	40/85	0.2270	62/85
nxCG@25	0.2255	34/85	0.2167	39/85	0.1826	54/85
nxCG@50	0.1801	28/85	0.1742	31/85	0.1466	47/85

Another observation is that structural conditions do not improve results which can be explained by the fact that only a restricted set of XML elements are treated when score aggregation is not used. Having results for strict quantisation could lead to further conclusions notably with regard to the run V2006CH0x3000foc whose configuration gave good results with the INEX'2005 data.

Table 3. Results for the subtask BestInContext

Task	BestInContext					
Metric: BEP-D						
Run	V2006CH01xp1bic		V2006CH06xp50bic		V2006CASCH06xp50bic	
	BEPD	rank	BEPD	rank	BEPD	rank
At 0.01	0.1175	30/77	0.1088	35/77	0.0015	75/77
At 0.1	0.1826	30/77	0.1702	38/77	0.0039	75/77
At 1.0	0.2958	30/77	0.2907	34/77	0.0079	75/77
At 10.0	0.4729	34/77	0.4832	30/77	0.0122	75/77
At 100.0	0.6430	39/77	0.6711	30/77	0.0185	75/77

A first observation is that configuring our method with weak score aggregation and with no query presence factor or configuring our method with score aggregation and with query presence factor seems to lead to close evaluations. Further investigations at the query level have to be carried out to compare the results of these two runs to determine the proportion of common ranked elements and eventually to consider a possible fusion strategy.

A second observation is related to the weak results given using the castitle part of the queries (i.e. run V2006CASCH06xp50bic). Introducing structural hints seems to move

Table 4. Results for the subtask AllInContext

Task	AllInContext					
Metric: generalized Precision/Recall						
Run	V2006CH01x1ric		V2006CH06x50ric		V2006CASCH06x50ric	
	MAgp	rank	MAgp	rank	MAgp	rank
	0.0441	50/56	0.0835	39/56	0.0887	37/56

away the selected elements from the relevant ones. A too high value fixed for the coefficients associated to structural treatment could be a reason.

The results for this subtask are weaker than for the other subtasks. Additional evaluations with respect to article level and element level seem to indicate that the difficulty seems to exist in the element retrieval rather than in the article retrieval. Further analysis has to be carried out to find explanations and to consider evolutions of our method.

5 Conclusions

Different changes have been introduced between the previous INEX'2005 round and the current INEX'2006 round. The changes occurred at different levels:

- The Wikipedia corpus has replaced the IEEE corpus introducing differences on corpus size, document contents and document structures,
- A new task called 'Best in Context' has been defined that asks systems to return one best entry per relevant article,
- There is no separate CAS task. Runs using topic 'titles' and runs using topic 'castitles' have been merged for evaluation. Furthermore, it was possible to make runs using other topic parts that title part or castitle part.

Participating to INEX this year had multiple objectives such as evaluating the impact of framework changes on our method effectiveness and to study the behaviour of different configurations of our method resulting from learning on previous INEX testbeds.

The results lead to mixed conclusions. The behaviour of the different configurations of our method using INEX'2006 data is similar to the behaviour of the same configurations when learning on INEX'2005 data. However, additional results show that other configurations of our method lead to better results on INEX'2006 data. A wider range of domains included in the INEX'2006 could explain this difference. Furthermore, the results show that a unique configuration of our method does not fit all the subtasks defined. A given configuration seems to be more adapted to a given subtask. This leads to consider a future study to determine how to configure our method to suit a given subtask. A first study [5] has been carried out on INEX'2005 data notably for the Thorough task. This study aims to identify how to configure our method according to different quantisations.

References

[1] Augé, J., Englmeier, K., Hubert, G., Mothe, J.: Classification automatique de textes basée sur des hiérarchies de concepts. Veille Stratégique Scientifique & Technologique (VSST'2001), Barcelona, 2001, pp. 291–300 (2001)

[2] Crouch, C.J., Khanna, S., Potnis, P., Doddapaneni, N.: The Dynamic Retrieval of XML ElementsAn Approach to Structured Retrieval Based on the Extended Vector Model. In: Fuhr, N., Lalmas, M., Malik, S., Kazai, G. (eds.) INEX 2005. LNCS, vol. 3977, pp. 268–281. Springer, Heidelberg (2006)

[3] Fuhr, N., Großjohann, K.: XIRQL: An XML query language based on information retrieval concepts. ACM Transactions on Information Systems (TOIS) 22(2), 313–356 (2004)

[4] Geva, S.: GPX – Gardens Point XML IR at INEX 2005. In: Fuhr, N., Lalmas, M., Malik, S., Kazai, G. (eds.) INEX 2005. LNCS, vol. 3977, pp. 240–253. Springer, Heidelberg (2006)

[5] Hubert, G., Mothe, J., Englmeier, K.: Tuning Search Engine to Fit XML Retrieval Scenario. 3rd International Conference on WEB Information Systems and Technologies (WEBIST 2007), Barcelona, pp. 228–233 (2007)

[6] Hubert, G.: XML Retrieval Based on Direct Contribution of Query Components. In: Fuhr, N., Lalmas, M., Malik, S., Kazai, G. (eds.) INEX 2005. LNCS, vol. 3977, pp. 172–186. Springer, Heidelberg (2006)

[7] Hubert, G.: A voting method for XML retrieval. In: Fuhr, N., Lalmas, M., Malik, S., Szlávik, Z. (eds.) INEX 2004. LNCS, vol. 3493, pp. 183–196. Springer, Heidelberg (2005)

[8] Larson, R.R.: Probabilistic Retrieval, Component Fusion and Blind Feedback for XML Retrieval. In: Fuhr, N., Lalmas, M., Malik, S., Kazai, G. (eds.) INEX 2005. LNCS, vol. 3977, pp. 225–239. Springer, Heidelberg (2006)

[9] Mass, Y., Mandelbrod, M.: Using the INEX Environment as a Test Bed for Various User Models for XML Retrieval. In: Fuhr, N., Lalmas, M., Malik, S., Kazai, G. (eds.) INEX 2005. LNCS, vol. 3977, pp. 187–195. Springer, Heidelberg (2006)

[10] Mihajlović, V., Ramírez, G., Westerveld, T., Hiemstra, D., Blok, H.E., de Vries, A.P.: TIJAH Scratches INEX 2005: Vague Element Selection, Image Search, Overlap, and Relevance Feedback. In: Fuhr, N., Lalmas, M., Malik, S., Kazai, G. (eds.) INEX 2005. LNCS, vol. 3977, pp. 72–87. Springer, Heidelberg (2006)

[11] Ogilvie, P., Callan, J.: Parameter Estimation for a Simple Hierarchical Generative Model for XML Retrieval. In: Fuhr, N., Lalmas, M., Malik, S., Kazai, G. (eds.) INEX 2005. LNCS, vol. 3977, pp. 211–224. Springer, Heidelberg (2006)

[12] Sigurbjörnsson, B., Kamps, J., de Rijke, M.: The Effect of Structured Queries and Selective Indexing on XML Retrieval. In: Fuhr, N., Lalmas, M., Malik, S., Kazai, G. (eds.) INEX 2005. LNCS, vol. 3977, pp. 104–118. Springer, Heidelberg (2006)

[13] Sauvagnat, K., Hlaoua, L., Boughanem, M.: XFIRM at INEX 2005: Ad-Hoc and Relevance Feedback Tracks. In: Fuhr, N., Lalmas, M., Malik, S., Kazai, G. (eds.) INEX 2005. LNCS, vol. 3977, pp. 88–103. Springer, Heidelberg (2006)

[14] Sauvagnat, K., Hubert, G., Boughanem, M., Mothe, J., IRIT,: at INEX, 2nd INitiative for the Evaluation of XML Retrieval, Dagstuhl, 2003, pp. 142–148 (2003)

[15] Vittaut, J.-N., Piwowarski, B., Gallinari, P.: An Algebra for Structured Queries in Bayesian Networks. In: Fuhr, N., Lalmas, M., Malik, S., Szlávik, Z. (eds.) INEX 2004. LNCS, vol. 3493, pp. 100–112. Springer, Heidelberg (2005)