Processing Heterogeneous Collections in XML Information Retrieval

Maria Izabel Menezes Azevedo and Klérisson Vinícius Ribeiro Paixão, and Diego Vinícius Castro Pereira

Department of Computer Science,
State University of Montes Claros, Montes Claros(MG), Brazil
mimaizabel@gmail.com, klerisson@hotmail.com, diegovcastro@yahoo.com.br

Abstract. Our model is based on the observation that the tags used in XML documents are semantically related to the content that they delimit. To evaluate the performance of our approach, we participated in the INEX 2004 heterogeneous track, along with 34 other institutions, from which only 5 groups, including us, submitted runs. In this paper we describe how the approach we used in INEX 2004 and 2005 processes heterogeneous collections without any mapping of DTDs.

1 Introduction

Our model [3] is based on the observation and theorical confirmation [1] that the tags of an XML [4] document are semantically related to the content that they delimit. We consider the structure of the XML standard as a new source of evidence, able to assist in the identification of the information contained in the documents without being essential to its identification.

According to this premise, formal aspects of XML, such as the DTD, are not used in our model. Our aim is to associate the XML tags with their content, based on statistical measures that are similar to those used in the standard vector space model, which relate the frequency of terms with the information of the document.

Thus, our model explores the diversity of tags of the XML documents, and its potential is evaluated in heterogeneous collections, where the structural diversity allows better linking between the semantics of tags and their content.

In this paper, we show how our model processes heterogeneous collections, presenting how each one of the subfactors are calculated and how they are placed in the standard vector space model to explore important aspects of the XML structure.

The remainder of this paper is organized in as follows. Section 2 presents related work. Section 3 is about the calculation of each subfactor in heterogeneous collections. Section 4 presents the results and Section 5 concludes the paper.

2 Related Work

In INEX 2004, according to Sauvagnat and Boughanem [12], the idea behind the heterogeneous track is that the information seeker is interested in semantically

N. Fuhr et al. (Eds.): INEX 2005, LNCS 3977, pp. 388–397, 2006.

meaningful answers irrespective of the structure of documents. Thus, this idea motivated their model which uses the relevance propagation method. This model is based on automatic indexing and introduces an interesting query processing technique that is able to process sub-queries that are logically linked. For this, the first step is to transform NEXI [13] topics into XFIRM queries. Then, this new query is decomposed into sub-queries. After each sub-query has been processed, the result of each one of them is propagated to generate the whole result of the query. However, mapping structural conditions from one DTD onto another was a problem, and to solve it they presented one DTD built manually by comparing the different DTDs.

Another relevant work was presented by [8], which continued to explore the approach of fusion to XML retrieval. This approach, in the heterogeneous track, was used to treat the different collections as separate databases with their own DTD, but these databases can be treated as a single database by the system. Another important work, described by Larson, was the configuration file which could specify subsets of tags to be used with the same meaning, for example `//p`, `//p1`, `//tf` for "paragraphs". We suppose Larson did not have problems with tags without semantic meaning such as Fld001, but he did not mention it.

In [2], an approach for creating a unified heterogeneous structure from heterogeneous data sources was presented. To build this unified conceptual model, first they identified groups of concepts that are semantically similar. To do this, they used an approach called WordNet, developed by Christine Fellbaum [5], which is able to detect similarity between "editor" and "edition", for example. Finally, to treat tags without semantic meaning it is necessary to capture the DTD comments preceding them and searches for the best cluster to put them.

All the participants of the heterogeneous track had difficulties to treat one document that was 217MB in size. To solve this problem, Larson [8] proposed to treat each of the main sub-elements as separate documents. Lehtonen [9] proposed to divide the file into fragments based on size. Although, large documents are problematic at "Het Track", they are not relevant problems of "Het Track" only, but to retrieval systems in general.

3 XML Factor Calculation

In our previous paper [3], we have defined how the standard vector space model [10] [11] was adapted to process an XML document whose relevance is given by:

$$\rho(Q, D) = \sum ti \in Q \cap D \frac{W_q(ti) * W_d(ti) * fxml(ti, e)}{Q * D} \quad (1)$$

where,

- *ti* is a term in the collection;
- $W_q(ti)$ is the weight of the query;
- $W_d(ti)$ is the weight of the document;
- *fxml(ti,e)* is the XML factor.

Next, we describe how the factor *fxml* is used to process heterogeneous collections and return elements of different structures without any mapping between those structures.

We consider the following characteristics of the XML standard:

- Its nested structure through the Nesting Factor *(fnh)*;
- The similarity between the query structure and the structure of the document, through the Structure Factor *(fstr)*;
- The semantic relation between terms and tags, through the Co-occurrence Factor *(focr)*.

Then, the factor *fxml* is given by:

$$fxml(ti, e) = fnh(ti, e) * fstr(ti, e) * focr(ti, e) \tag{2}$$

where,

- *ti* is a term in the collection;
- *e* is an element in the collection.

3.1 The Nesting Factor

The nesting factor expresses the importance of terms considering their positions in the XML tree. As the augmentation factor, this factor reduces the term contribution according to the distance of the elements in XML tree. The factor proposed does not have a defined value. Its value is inversely proportional to the distance between the level of the element that contains the term and its ancestor, whose relevance is calculated. It is given by:

$$fnh(ti, e) = \frac{1}{(1 + nl)} \tag{3}$$

where,

- *nl* is the number of levels from element *e* to its sub-element that contains the term *ti*.

The nesting factor can vary between the following values:

- *fnh(ti,e) = 1*, for terms directly in elements *e*;
- *fnh(ti,e) = 1/nd*, where *nd* is the depth of the XML tree.

In Figure 1, we have *fnh(*`computer,fm`*) = 1/(1+3)*. This factor reduces the contribution of one term for the relevance of the elements more distant (upwards) in the XML tree. It compensates the high frequency of a term in upwards elements caused by the nesting of the XML structure. Without this consideration the upwards elements tend to occupy the first positions in the ranking showed to the user.

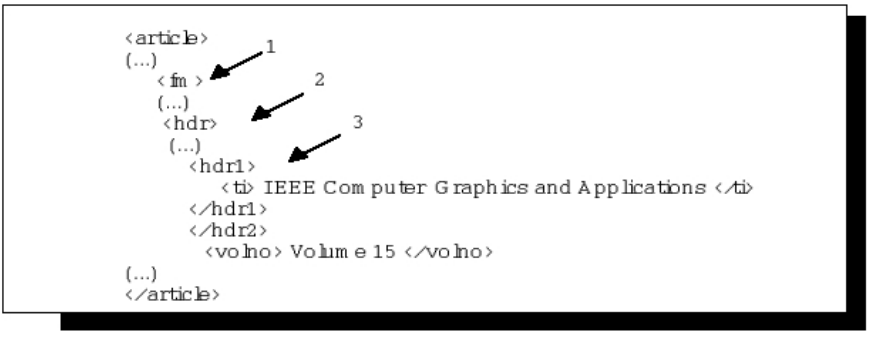


Fig. 1. The Nesting Factor

3.2 The Structure Factor

The Structure Factor expresses how a query with structural constraints (CAS) [7] is satisfied by the context of the determined element. This factor values a context that better satisfies structural constraints present in the query. Mathematically, it is given by the relation between structural constraints in the query and the structure of the element. It is given by:

$$fstr(ti, e) = \frac{(common\_markups + 1)}{(nr\_qmarkups + 1)} \tag{4}$$

where,

- *common_markups* is the number of tags presented in the query structural constraints and also in the context of element *e* that contains *ti*;
- *nr_markups* is the number of tags in the query structural constraints.

It can vary from:

- *fstr(ti,e) = 1/(nr_qmarkups+1)*, when no structural constraints appears in the context of *ti*;
- *fstr(ti,e) = 1*, when all query's structural constraints tags appears in the context of *ti*.

For example, for the NEXI query in Figure 2 processed on the heterogeneous collection, also shown in Figure 2, we obtain the following results:

For the element `/artigo/pessoa/nome` in the first article:

- *nr_qmarkups = 3;*
- *common_markups = 2;*

and

- *fstr(ti, e) = (2+1)/(3+1) = 3/4.*

For the element `/article/person/name` in the second article:

- *nr_markups = 3;*
- *common_markups = 0;*

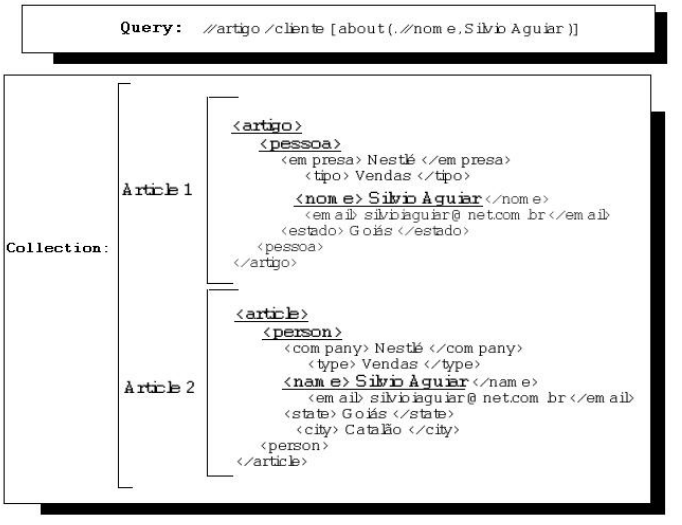


Fig. 2. The Structure Factor

and

– *fstr(ti,e) = (0+1)/(3+1) = 1/4.*

This factor gives more value to the element `/artigo/pessoa/nome` in the first article, where *fstr* is equal to *3/4*, and whose context better meets the structural constraints of the query. However, it does allow that the element `/article/person/name` in the second article, where *fstr* is equal to *1/4*, is also returned to the user, demonstrating the model's application on heterogeneous collections.

This is important for CAS queries, where the user specifies the elements that will better fit his or her information need. For CO [7] queries, *fstr* will always be equal to 1 because of the following:

– *nr_qmarkups = 0* (CO queries do not have any structural constraints);
– *common_markups = 0* (there are no common tags between documents and query);

and

– *fstr (ti,e) = 1.*

Consequentially, in this case the factor *fstr* does not influence the relevance equation.

3.3 The Co-occurrence Factor

The last factor, Co-occurrence Factor, expresses the semantic relation between tags and their contents. To express mathematically this semantic relation, we

applied the same principle that is used in standard vector space model to relate terms and documents: the higher the frequency of a term in one document, the greater the semantical relation between them. The value of this semantical relation for a particular query depends on the frequency of terms and tags in the collection, too. This factor is calculated as follows.

$$focr(ti, e) = cf(ti, e) * idf(ti, e) * N * icf(e) \quad (5)$$

where,

- *cf(ti,e)* is the number of times the tag of element e, denoted by m, delimits a textual content containing term *ti*. In other words, number of co-occurrences of term *ti* and tag m in the collection;
- *idf(ti,e)* is the inverse of the number of elements e that contain *ti*.

So, $cf(ti, e) * idf(ti, e)$, is the reason between the number of times term *ti* appears with m for the number of the elements containing *ti* in the collection.

- *icf(e)* is the inverse of the number of times markup m appears in the collection.
- N is the total number of elements in the collection;

and

- $icf(e) * N$ express the popularity of tags m in the collection.

For example, for the NEXI query of Figure 3 processed on the heterogeneous collection shown in Figure 3, we have the following results:

For the element `/artigo/pessoa/estado` in the first article:

- *cf(Paulo, estado) = 1;*
- *idf(Paulo) = 1/4;*
- *icf(estado) = 1/3;*
- *N = 16;*
- *focr(Paulo, estado) = (1 * 1/4) * (1/3 * 16) = 1.333*

For the element `/artigo/pessoa/nome` in the second article:

- *cf(Paulo, nome) = 2;*
- *idf(Paulo) = 1/4;*
- *icf(nome) = 1/3;*
- *N = 16;*
- *focr(Paulo, nome) = (2 * 1/4) * (1/3 * 16) = 2.666.*

For the element `/article/person/name` in the last article:

- *cf(Paulo, name) = 1;*
- *idf(Paulo) = 1/4;*
- *icf(name) = 1/1;*
- *N = 16;*
- *focr(Paulo, name) = (1 * 1/4) * (1/1 * 16) = 4.0.*

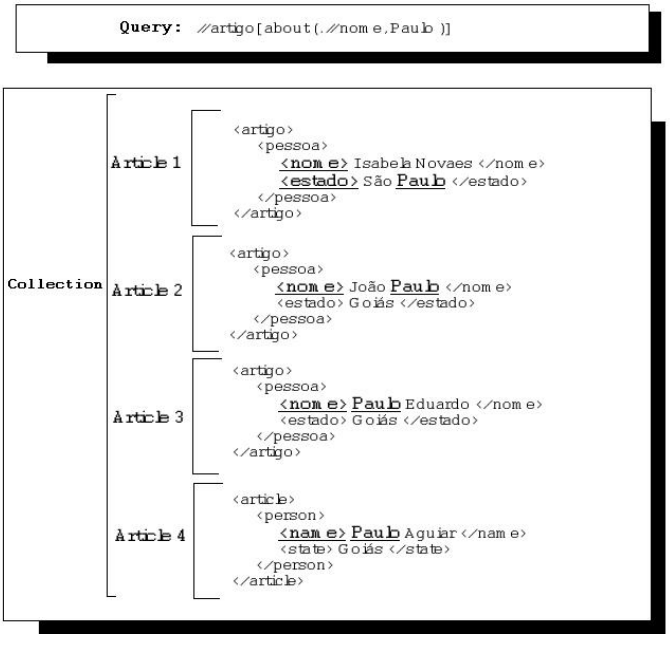


Fig. 3. The Co-occurrence Factor

The fact of `/article/person/name` and `/artigo/pessoa/nome` be returned to the user confirms that our model can deal with heterogeneous collections. The element `/article/person/name` will be presented to the user for having been valued for its rarity within the collection. The element `/artigo/pessoa/estado` will be presented to the user although it does not meet the query but obtains the lower value of co-occurrence factor.

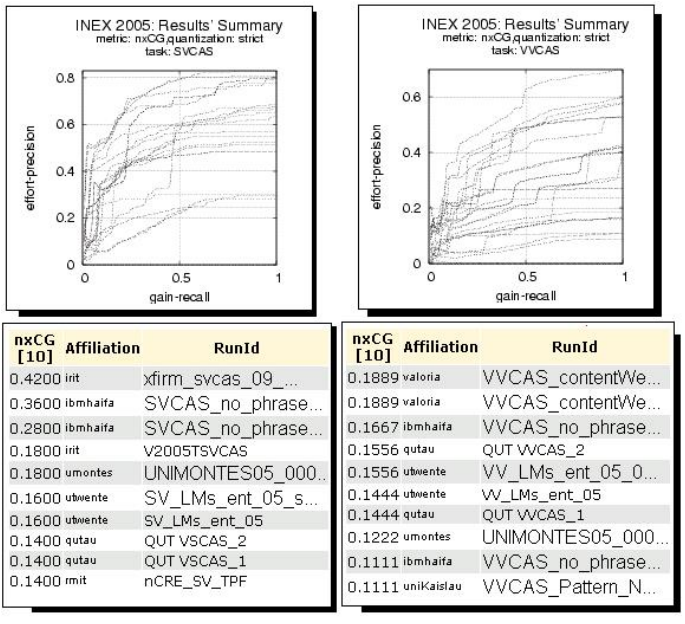
The co-occurrence factor values the co-occurrence of terms and tags, considering the popularity of tags. With this factor we intend to explore the characteristic of XML originating from its definition: the presence of tags that describe its contents.

For the effectiveness of our model, it is important that it has a narrow semantic relation between terms and tags. This semantic relation will be easier in heterogeneous collections because they present a bigger structure diversity.

To conclude, the XML Factor *(fxml)* explores characteristics of XML by looking for the semantics of terms and information behind words.

4 Results

We submitted runs to the INEX heterogeneous track, but as the assessments were not concluded yet at the time of writing, we have no Recall/Precision curves to show here. It follows an answer to a query containing elements from



nxCG [10]	Affiliation	RunId
0.4200	irit	xfirm_svcas_09_...
0.3600	ibmhaifa	SVCAS_no_phrase...
0.2800	ibmhaifa	SVCAS_no_phrase...
0.1800	irit	V2005TSVCAS
0.1800	umontes	UNIMONTES05_000...
0.1600	utwente	SV_LMs_ent_05_s...
0.1600	utwente	SV_LMs_ent_05
0.1400	qutau	QUT VSCAS_2
0.1400	qutau	QUT VSCAS_1
0.1400	rmit	nCRE_SV_TPF

nxCG [10]	Affiliation	RunId
0.1889	valoria	VVCAS_contentWe...
0.1889	valoria	VVCAS_contentWe...
0.1667	ibmhaifa	VVCAS_no_phrase...
0.1556	qutau	QUT VVCAS_2
0.1556	utwente	VV_LMs_ent_05_0...
0.1444	utwente	VV_LMs_ent_05
0.1444	qutau	QUT VVCAS_1
0.1222	umontes	UNIMONTES05_000...
0.1111	ibmhaifa	VVCAS_no_phrase...
0.1111	uniKaislau	VVCAS_Pattern_N...

Fig. 4. Adhoc Results for SVCAS and VVCAS sub tasks

many sub-collections, confirming that our model can deal with different DTDs. For query:

```
//article[about(.//author,Nivio Ziviani)]
```

we get the following answer:

```
<topictopic-id=‘‘2’’> ...
<result>
    <subcollection name=‘‘ieee’’/>
    <file>co/2000/ry037</file>
    <path>/article[1]/fm[1]/au[1]</path>
    <rank> 3</rank>
</result> ...
<result>
    <subcollection name=‘‘dblp’’ />
    <file>dblp</file>
    <path>/dblp[1]/article[177271]/author[4]</path>
    <rank>6</rank>
</result> ...
<result>
    <subcollection name=‘‘CompuScience’’ />
    <file>exp-dxf1.xml.UTF-8</file>
    <path>/bibliography[1]/article[23]/author[1]</path>
```

```
    <rank> 30</rank>
</result>
...
<result>
    <subcollectionname=‘‘hcibib’’ />
    <file>hcibib</file>
    <path>/file[1]/entry[229]/article[1]/author[1]</path>
    <rank>139</rank>
</result>
```

At INEX 2005, we submitted runs to the ad-hoc track. For SVCAS and VVCAS [6] topics, our results were in the top ten as shown in Figure 4.

Analysing these results and observing the concept of both topics, we can conclude that our system worked better for CAS topics, where support elements have been interpreted vaguely. It is coherent with our information retrieval view. We have developed an approach where an element is relevant if it satisfies the information need, irrespective of the structural constraints. We get better results when target element constraints are strictly satisfied, showing that the structure factor (fstr), proposed in our model, can improve performance.

5 Conclusion

In INEX 2005, we get better results than at INEX 2004, showing that our research is going in a correct direction. The ability of our system to deal with a heterogeneous collection and its results when target element constraints are strictly satisfied shown that the structure factor (fstr) proposed in our model can improve performance. For CO queries, specially the CO+S ones, we do get the worst results, demanding further investigation.

For next year, we intend to participate in the heterogeneous track so that we can really evaluate the co-occurrence factor (fcoo) and conclude if tags of XML structure can be used to improve performance of search engines.

References

1. S. Abiteboul, P. Buneman and D. Suciu.: Data on the Web - From Relations to Semistructured Data in XML. Mogan Kaufmann Publishers, San Francisco, California, (2000) 27–50.
2. S. Abiteboul, I. Manolescu, B. Nguyen, N. Preda.: A Test Plataform for the INEX Heterogeneous Track. INEX (2004) LNCS
3. M. Azevedo, L.Pantuza e N. Ziviane.: A Universal Model for XML Information Retrieval. INEX (2004) LNCS 3493 311–321 (2005).
4. T. Bray, J. Paoli, C. M. Sperberg-McQueen and E. Maler.: Extensible Markup Language (XML) 1.0. 2nd ed. http://www.w3.org/TR/REC-xml, Cct 2000. W3C Recommendation 6 October (2000).
5. C. Fellbaum.: *WordNet*: An Electronic Lexical Database. MIT Press. (1998).

6. S. Geva, M. Lalmas, B. Larsen, S. Malink, B. Sigurbjrnsson and A. Trotman.: INEX 2005 Guidelines for Topic Development. In INEX 2005 Workshop Pre-Proceedings,(2005) pp.375.
7. G. Kazai, M. Lalmas and S. Malik.: INEX'03 Guidelines for Topic Development. In INEX 2003 Workshop Proceedings, Duisburg, 2003 pg. 153-154.
8. R. R. Larson.: Cheshire II at INEX '04: Fusion and Feedback for the Adhoc and Heterogeneous Tracks. INEX (2004) LNCS 3493 322–336.
9. M. Lehtonen.: Extirp 2004: Towards Heterogeneity. INEX (2004) LNCS 3493 372–381.
10. B. Ribeiro-Neto e R. Baeza-Yates.: Modern Information Retrieval. Addison Wesley. (1999) pp. 27–30.
11. G. Salton e M. E. Lesk.: Computer evaluation of indexing and text processing. Journal of the ACM. 15(1) (1968) 8–36.
12. K. Sauvagnat, M. Boughanem.: Using a relevance propagation method for Adhoc and Heterogeneus tracks in INEX 2004. INEX (2004) LNCS 3493 337–348.
13. A. Trotman and B. Sigurbjornsson.: Narrowed extended xpath i. In In INEX 2004 Workshop Proceedings,(2004) pp.16.