BUAP: A First Approach to the Data-Centric Track of INEX 2010*

Darnes Vilariño, David Pinto, Carlos Balderas, Mireya Tovar, and Saul León

Facultad de Ciencias de la Computación
Benemérita Universidad Autónoma de Puebla, México
{darnes,dpinto,mtovar}@cs.buap.mx, charlie_kanon@hotmail.com,
saul.ls@live.com

Abstract. In this paper we present the results of the evaluation of an information retrieval system constructed in the Faculty of Computer Science, BUAP. This system was used in the Data-Centric track of the Initiative for the Evaluation of XML retrieval (INEX 2010). This track is focused on the extensive use of a very rich structure of the documents beyond the content. We have considered topics (queries) in two variants: Content Only (CO) and Content And Structure (CAS) of the information need. The obtained results are shown and compared with those presented by other teams in the competition.

1 Introduction

Current approaches proposed for keyword search on XML data can be categorized into two broad classes: one for document-centric XML, where the structure is simple and long text fields predominate; the other for Data-Centric XML, where the structure is very rich and carries important information about objects and their relationships [1]. In previous years, INEX focused on comparing different retrieval approaches for document-centric XML, while most research work on Data-Centric XML retrieval cannot make use of such a standard evaluation methodology. The new Data-Centric track proposed at INEX 2010 aims to provide a common forum for researchers or users to compare different retrieval techniques on Data-Centric XML, thus promoting the research work in this field [2].

Compared to traditional information retrieval, where whole documents are usually indexed and retrieved as single complete units, information retrieval from XML documents creates additional retrieval challenges.

Until recently, the need for accessing the XML content has been addressed diferently by the database (DB) and the information retrieval (IR) research communities. The DB community has focussed on developing query languages and eficient evaluation algorithms used primarily for Data-Centric XML documents. On the other hand, the IR community has focussed on document-centric XML documents

* This work has been partially supported by the CONACYT project #106625, VIEP # VIAD-ING11-I, as well as by the PROMEP/103.5/09/4213 grant.

S. Geva et al. (Eds.): INEX 2010, LNCS 6932, pp. 219–226, 2011.

by developing and evaluating techniques for ranked element retrieval. Recent research trends show that each community is willing to adopt the well-established techniques developed by the other to efectively retrieve XML content [3].

The Data-Centric track uses the IMDB data collection newly built from the following website: http://www.imdb.com. It consists of information about more than 1,590,000 movies and people involved in movies, e.g. actors/actresses, directors, producers and so on. Each document is richly structured. For example, each movie has title, rating, directors, actors, plot, keywords, genres, release dates, trivia, etc.; and each person has name, birth date, biography, filmography, and so on.

The Data-Centric track aims to investigate techniques for finding information by using queries considering content and structure. Participating groups have contributed to topic development and evaluation, which will then allow them to compare the effectiveness of their XML retrieval techniques for the Data-Centric task. This will lead to the development of a test collection that will allow participating groups to undertake future comparative experiments.

2 Description of the System

In this section we describe how we have indexed the corpus provided by the task organizers. Moreover, we present the algorithms developed for tackling the problem of searching information based on structure and content.

The original XML file has been previously processed in order to eliminate stopwords and punctuation symbols. Moreover, we have transformed the original hierarchical structure given by XML tags to a similar representation which may be easily analyzed by our parser.

For the presented approach we have used an inverted index tree in order to store the XML templates of the corpus. The posting list contains the reference of the document (document ID) and the frequency of the indexed term in the given context (according to the XML tag).

With respect to the dictionary of the inverted index, we have considered to include both, the term and the XML tag (the last one in the hierarchy). In Figure 1, we show an example of the inverted index (pay special attention to the dictionary). The aim was to be able to find the correct position of each term in the XML hierarchy and, therefore, to be able to retrieve those parts of the XML file containing the correct answer of a given query. In this way, the inverted index allows to store the same term which occurs in different contexts. We assumed that the last XML tag would be enough for identifying the full path in which the term occurs, however, it would be better to use all the hierarchy in the dictionary. Further experiments would verify this issue.

In the following subsection we present the algorithms developed for indexing and searching information based on content and structure.

2.1 Data Processing

Before describing the indexing techniques used, we first describe the way we have processed the data provided for the competition. We have cleaned the XML files

```
title hannibal 40     : 981731:1 994171:1  78811:1 [...] 1161611:1 [...]

character lecter 440 : 959641:1 959947:1  1161611:1 969446:1 [...]

name hopkins 3068     : 1154469:1 1154769:2 1154810:1 [...] 1161611:1 [...]

name anthony 31873    : 943773:1 [...] 944137:2 1161611:1 1224420:3 [...]

director scott 4771   : 1157203:1 1157761:1 1157773:1 [...] 1161611:1 [...]

director ridley 62    : 1289515:1 1011543:1 1011932:1 [...] 1161611:1 [...]

writer harris 2114    : 1120749:1 1121040:1 1121294:1 [...] 1161611:1 [...]

writer thomas 7333    : 115985:1 115986:1 [...] 1161611:1 1161616:2 [...]

        :                                             :
```

Fig. 1. Example of the type of inverted index used in the experiments

in order to obtain an easy way of identifying the XML tag for each data. For this purpose, as we previously mentioned, we have traduced the original hierarchical structure given by XML tags to a similar representation which may be easily analyzed by our parser. Thereafter, we have created five different inverted indexes, for the each one of the following categories: actors, directors, movies, producers and others. The inverted index was created as mentioned in the previous section.

Once the dataset was indexed we may be able to respond to a given query. In this case, we have also processed the query by identifying the corresponding logical operators (AND, OR). Let us consider the query presented in Figure 2, which is then traduced to the sentence shown in Figure 3. The first column is the topic or query ID; the second column is the number associated to the ct_no tag; the third column indicates the number of different categories that will be processed, in this case, we are considering only one category: movies.

In the competition we submitted two runs. The first one uses the complete data of each record (word n-gram), whereas the second approach considered to split the data, that corresponds to each content, into unigrams, with the goal of being more specific in the search process. However, as will be seen in the experimental results section, both approaches perform similar. An example topic showing the second approach is given in Figure 4, whereas its traduced version is given in Figure 5.

In order to obtain the list of candidate documents for each topic, we have calculated the similarity score between the topic and each corpus document as shown in Eq. (1) [4], which was implemented as presented in Algorithm 1.

```
<topic id="2010001" ct_no="3">
  <title>Yimou Zhang 2010 2009</title>
  <castitle>//movie[about(.//director, "Yimou Zhang") and
                    (about(.//releasedate, 2010) or
                     about(.//releasedate, 2009))]
  </castitle>
  <description>I want to know the latest movies directed by Yimou Zhang.
  </description>
  <narrative>I am interested in all movies directed by Yimou Zhang, and
             I want to learn the latest movies he directed.
  </narrative>
</topic>
```

Fig. 2. An example of a given query (topic)

```
2010001 3 1 //movie//director yimou zhang and //movie//releasedate 2010
            or //movie//releasedate 2009
```

Fig. 3. Representation of the topic

```
<topic id="2010025" ct_no="19">
  <title>tom hanks steven spielberg</title>
  <castitle>//movie[about(., tom hanks steven spielberg)]</castitle>
  <description>movies where both tom hanks and steven spielberg worked
               together
   </description>
  <narrative>The user wants all movies where Tom Hanks and Steven
             Spielberg have worked together (as actors, producers,
             directors or writers). A relevant movie is a movie
             where both have worked together.
  </narrative>
</topic>
```

Fig. 4. An example of another topic

```
2010025 19 1 //movie tom and //movie hanks and //movie steven and
             //movie spielberg
```

Fig. 5. The representation of a topic by splitting the data

$$\mathrm{SIM}(q,d) = \sum_{c_k \in B} \sum_{c_l \in B} CR(c_k, c_l) \sum_{t \in V} weight(q,t,c_k) \frac{weight(d,t,c_l)}{\sqrt{\sum_{c \in B, t \in V} weight(d,t,c)^2}} \tag{1}$$

where the CR function is calculated as shown in Eq. (2), V is the vocabulary of non-structural terms; B is the set of all XML contexts; and $weight(q,t,c)$ and $weight(d,t,c)$ are the weights of term t in XML context c in query q and document d, respectively (as shown in Eq. (3)).

$$\mathrm{CR}(c_q, c_d) = \begin{cases} \frac{1+|c_q|}{1+|c_d|} & \text{if } c_q \text{ matches } c_d \\ 0 & \text{otherwise} \end{cases} \tag{2}$$

where c_q and c_d are the number of nodes in the query path and document path.

$$weight(d, t, c) = idf_t * wf_{t,d} \tag{3}$$

where idf_t is the inverse document frequency of term t, and $wf_{t,d}$ is the frequency ot term t in document d.

Algorithm 1. Scoring of documents given a topic q

```
Input: $q$, $B$, $V$, $N$ : Number of documents, $normalizer$
Output: $score$
for $n = 1$ to $N$ do
    $score[n] = 0$
    foreach $\langle c_q, t \rangle \in q$ do
        $w_q = \mathrm{weight}(q, t, c_q)$
        foreach $c \in B$ do
            if $CR(c_q, c) > 0$ then
                $postings = \mathrm{GetPostings}(c, t)$
                foreach $posting \in postings$ do
                    $x = CR(c_q, c) * w_q * \mathrm{PostingWeight}(posting)$
                    $score[docID(posting)] += x$
                end
            end
        end
    end
end
for $n = 1$ to $N$ do
    score[n] = score[n]/$normalizer$[n]
end
return $score$
```

2.2 Experimental Results

We have evaluated 25 topics with the corpus provided by the competition organizers. This dataset is made up of 1,594,513 movies, 1,872,492 actors, 129,137 directors, 178,117 producers and, finally, 643,843 files categorized as others.

As it was mentioned before, we submitted two runs which we have named: “FCC-BUAP-R1” and “FCC-BUAP-R2”. The former uses the complete data of each record (n-gram), whereas the latter split the words contained in the query by unigrams. The obtained results, when evaluating the task as focused retrieval (MAgP measure) are presented in Table 1 and in Figure 6.

As may be seen, we have obtained a low performance, which we consider is derived of the fact of using only one tag for identifying each indexed term. We

Table 1. Evaluation measured as focused retrieval (MAgP)

#	MAgP	Institute	Run
1	0.24910409	University of Otago	OTAGO-2010-DC-BM25
2	0.24585548	Universitat Pompeu Fabra	UPFL15TMI
3	0.24337897	Universitat Pompeu Fabra	UPFL15TMImov
4	0.18113477	Kasetsart University	NULL
5	0.15617634	University of Otago	OTAGO-2010-DC-DIVERGENCE
6	0.06517544	INDIAN STATISTICAL INSTITUTE	ISI_fdbk_em_10
7	0.0587039	INDIAN STATISTICAL INSTITUTE	ISI_nofdbk_article_.4.-1
8	0.04490731	INDIAN STATISTICAL INSTITUTE	ISI_nofdbk_article_.4.0
9	0.04426635	INDIAN STATISTICAL INSTITUTE	ISI_fdbk_10
10	0.04091211	B. Univ. Autonoma de Puebla	FCC-BUAP-R1
11	0.04037697	B. Univ. Autonoma de Puebla	FCC-BUAP-R2
12	0.03788804	INDIAN STATISTICAL INSTITUTE	ISI_nofdbk_article_.6.0
13	0.03407941	INDIAN STATISTICAL INSTITUTE	ISI_nofdbk_article_.4.1
14	0.02931703	INDIAN STATISTICAL INSTITUTE	ISI_nofdbk_article_.2.0

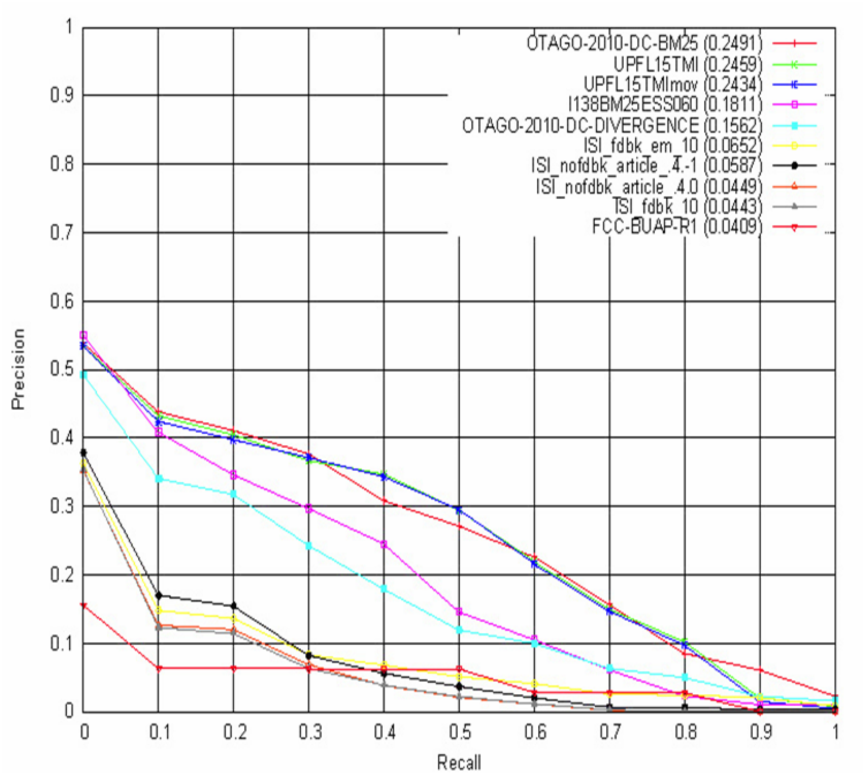

Fig. 6. Evaluation measured as focused retrieval (MAgP)

assumed that the last XML tag in each context would be enough for identifying the complete path in which the term occurs, however, it would be better to use the complete hierarchy in the dictionary. Once the gold standard is being released, we are considering to carry out more experiments in order to verify this issue.

The evaluation of the different runs at the competition, measured as document retrieval, may be found in Table 2.

Table 2. Evaluation measured as document retrieval (whole document retrieval)

#	MAP	Institute	Run
1	0.5046	SEECS, Peking University	NULL
2	0.5046	SEECS, Peking University	NULL
3	0.3687	Universitat Pompeu Fabra	UPFL15TMI
4	0.3542	Universitat Pompeu Fabra	UPFL15TMImov
5	0.3397	University of Otago	OTAGO-2010-DC-BM25
6	0.2961	Universitat Pompeu Fabra	UPFL15Tall
7	0.2829	Universidade Federal do Amazonas	ufam2010Run2
8	0.2822	Universitat Pompeu Fabra	UPFL45Tall
9	0.2537	Universidade Federal do Amazonas	ufam2010Run1
10	0.2512	Universidade Federal do Amazonas	ufam2010Run5
11	0.2263	Universidade Federal do Amazonas	ufam2010Run3
12	0.2263	Universidade Federal do Amazonas	ufam2010Run4
13	0.2263	Universidade Federal do Amazonas	ufam2010Run5
14	0.2103	University of Otago	OTAGO-2010-DC-DIVERGENCE
15	0.2044	Kasetsart University	NULL
16	0.1983	Universitat Pompeu Fabra	UPFL15Tmovie
17	0.1807	INDIAN STATISTICAL INSTITUTE	ISI_elts.0
18	0.18	INDIAN STATISTICAL INSTITUTE	ISI_elts.1
19	0.1783	INDIAN STATISTICAL INSTITUTE	ISI_elts.-1
20	0.1578	Universitat Pompeu Fabra	UPFL45Tmovie
21	0.1126	INDIAN STATISTICAL INSTITUTE	ISI_fdbk_em_10
22	0.0888	INDIAN STATISTICAL INSTITUTE	ISI_nofdbk_article_.4.-1
23	0.0674	INDIAN STATISTICAL INSTITUTE	ISI_nofdbk_article_.4.0
24	0.0672	INDIAN STATISTICAL INSTITUTE	ISI_fdbk_10
25	0.0602	B. Univ. Autonoma de Puebla	FCC-BUAP-R2
26	0.0581	INDIAN STATISTICAL INSTITUTE	ISI_nofdbk_article_.6.0
27	0.0544	B. Univ. Autonoma de Puebla	FCC-BUAP-R1
28	0.0507	INDIAN STATISTICAL INSTITUTE	ISI_nofdbk_article_.4.1
29	0.0424	INDIAN STATISTICAL INSTITUTE	ISI_nofdbk_article_.2.0

3 Conclusions

In this paper we have presented details about the implementation of an information retrieval system which was used to evaluate the task of focused retrieval of XML documents, in particular, in the Data-Centric track of the Initiative for the Evaluation of XML retrieval (INEX 2010).

We presented an indexing method based on an inverted index with XML tags embedded. For each category (movies, actors, producers, directors and others), we constructed an independent inverted index. The dictionary of the index considered both, the category and the indexed term which we assumed to be sufficient to correctly identify the specific part of the XML file associated to the topic.

Based on the low scores obtained, we may conclude that a more detailed description in the dictionary (including more tags of the XML hierarchy) is needed in order to improve the precision of the information retrieval system presented.

References

1. Wang, Q., Li, Q., Wang, S., Du, X.: Exploiting semantic tags in XML retrieval. In: Geva, S., Kamps, J., Trotman, A. (eds.) INEX 2009. LNCS, vol. 6203, pp. 133–144. Springer, Heidelberg (2010)
2. Wang, Q., Trotman, A.: Task description of INEX 2010 Data-Centric track. In: Proc. of INEX 2010 (2010)
3. Amer-Yahia, S., Curtmola, E., Deutsch, A.: Flexible and efficient XML search with complex full-text predicates. In: Proceedings of the 2006 ACM SIGMOD International Conference on Management of Data, pp. 575–586. ACM, New York (2006)
4. Manning, C.D., Raghavan, P., Schütze, H.: Introduction to Information Retrieval. Cambridge University Press, Cambridge (2009)