Vector Space Model for Legal XML Retrieval

Sorina CORNOIU
Consiuliul Legislativ, Calea 13 Septembrie nr. 1-3
Sector 5 Bucuresti

Abstract—**In recent years, W3C's XML (eXtensible Mark-up Language) has been accepted as a major means for efficient data management and exchange. The use of XML ranges over information formatting and storage, database information interchange, data filtering, as well as web services interaction. Due to the ever-increasing web exploitation of XML, an efficient approach to compare XML-based documents becomes crucial in information retrieval (IR) [7]. Legal information is often accessible via portal web sites. Legal documents typically combine structured and unstructured information, the former being tagged with markup languages such as XML (Extensible Markup Language) [1]. In this paper, I propose using a vector space model for legal XML retrieval.**

Index Terms—**computer science, information retrieval, knowledge based systems, knowledge representation, text processing**

I. INTRODUCTION

In the legal information retrieval systems, the information is usually searched by means of a full text search, every term in the texts of the documents can function as a search key [1]. In the databases the legal documents are thus indexed with the terms that occur in their natural language texts and with extra descriptive data called metadata. There are various ways to improve the search technology for accessing legal documents. Legal documents typically combine structured and unstructured information, the former, for instance, referring to common document architectures, reference structures and metadata information the latter involving the natural language texts. The structured information is increasingly tagged with markup languages such as XML (Extensible Markup Language) [1].

A typical ranking model for Information Retrieval (IR) is the vector space model where documents and queries are both represented as vectors in a space where each dimension represents a distinct indexing unit . I propose to extend the vector space model so as to be able to compare legal XML fragments and legal XML documents as objects of the same nature, using the context resemblance.

II. XML LEGAL DOCUMENTS

Data-centric documents have a regular and strict structure, and the content is usually not mixed with large stretches of unstructured information such as free text. This is the type of information usually stored in a relational or object-oriented database [1]. Document-centric documents are characterized by a less regular structure, often contain considerably large text fragments apart from the structured content. The documents of this latter category might not strictly adhere to a DTD (Document Type Definition) or XML schema, or possibly the DTD or schema might not have been specified at all. Furthermore users of the documents of this latter category will generally not be interested in retrieving data [1]. Legislation typically involves structured information including the division of a legal documents in for instance titles, chapters, sections and articles, and the typical metadata (e.g., indication of the date of enactment, the area of applicability and references to other statutes) that are assigned to the doucments or its parts [3]. Additionally, legislation contains large parts of unstructured information found in the natural language texts.

The structured information is increasingly tagged with markup languages such as XML (Extensible Markup Language) [3] . The use of the document structure allows generating a more precise answer to an information query. Instead of returning the complete document as the answer, a structural element or several elements are given [1]. Instead of returning the complete document as the answer, a structural element or several elements are given [3]. Legal documents that are marked up with XML tags can be considered as an example of document-centric objects. Legal XML documents represent hierarchically structured information and can be modeled as Ordered Labeled Trees (OLTs) .

The Community Official Journal texts, which constitute our working base, are made of many types of texts grouped in two main categories: legislation, information and notices. In this paper, we focus on regulations, directives, decisions and recommendations regardless of their category [4].

In [2],[10] we can find recommendations and legislative techniques for structuring the document. Community acts are generally drafted according to a standard structure (Fig. 1). The 'Title' comprises all the information in the heading of the act which serves to identify it. It may be followed by certain technical data (reference to the authentic language version, relevance for the EEA, serial number) which are inserted, where appropriate, between the title proper and the preamble [2]. 'Preamble' means everything between the title and the enacting terms of the act, namely the citations, the recitals and the solemn forms which precede and follow them [2].

Citations: at the beginning of the preamble, they indicate the legal basis of the act, the proposals, recommendations, initiatives, drafts... that must be obtained, and certain opinions and other non-mandatory procedural steps. Citations are generally introduced by the dedicated expression 'Having regard to" or „Acting in accordance with" [4].

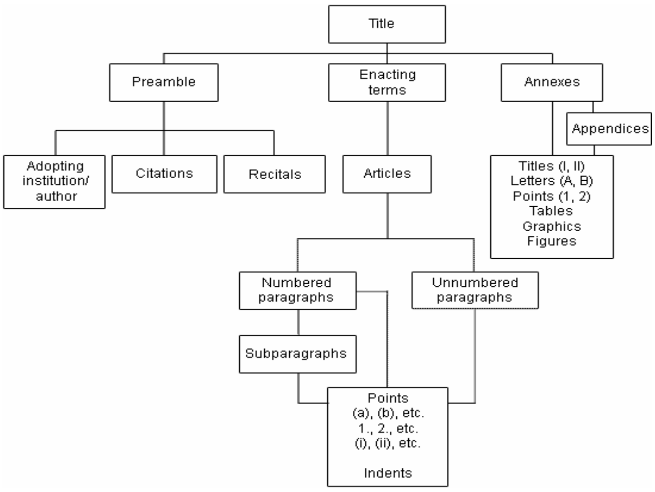

Figure 1. Basic structure of legislative acts.

Recitals: are the parts of the act containing the statement of reasons for the act; they are placed between the citations and the enacting terms. Recitals are introduced by the word 'Whereas : " and continue with numbered points comprising one or more complete sentences [4] . The 'enacting terms' are the legislative part of the act. They are composed of articles, which may be grouped in titles, chapters and sections, and may be accompanied by annexes [4] . „Annex" : where necessary, begins by the heading ''annex'' and is spread out until the end of document. In case where many annexes are necessary, each annex has a heading like the one cited above and is numbered [4] . Let's consider the next juridical document from EurLex : D0(32006D0191) - Commission Decision of 1 March 2006 declaring operational the Regional Advisory Council for the Baltic Sea under the common fisheries policy. In this case, we have the followind annotations:

```
  <act>
      <ti>
  <p>COMMISSION DECISION</p><p>of 1 March 2006</p>
  <p>declaring operational the Regional Advisory
Council for the Baltic Sea under the common
fisheries policy</p><p>(2006/191/EC)</p>
       </ti>
      <pr>
  <pr.init>THE COMMISSION OF THE EUROPEAN
COMMUNITIES,</pr.init>
  <pr.cit>
   <cit>Having regard to the Treaty establishing
the European Community,</cit>
  <cit>Having regard to Council Decision
2004/585/EC of 19 July 2004 establishing Regional
Advisory Councils under the common fisheries
policy (1), and in particular Article 3(3)
thereof,</cit>
  <cit>Having regard to the recommendation
transmitted by Denmark on 13 December 2005 on
behalf of Denmark, Germany, Estonia, Latvia,
Lithuania, Poland, Finland and Sweden,</cit>
       </pr.cit>
  <pr.rec>Whereas:
```

<rec>(1) Council Regulation (EC) No 2371/2002 of 20 December 2002 on the conservation and sustainable exploitation of fisheries resources under the common fisheries policy (2) and Decision 2004/585/EC provide the framework for the establishment and operation of Regional Advisory Councils.</rec>

<rec>(2) Article 2 of Decision 2004/585/EC establishes a Regional Advisory Council to cover the Baltic Sea in International Council for the Exploration of the Seas (ICES) areas IIIb, IIIc and IIId as defined in Council Regulation (EEC) No 3880/91 (3).</rec>

<rec>(3) In accordance with Article 3(1) of Decision 2004/585/EC, representatives of the fisheries sector and other interests groups submitted a request concerning the operation of that Regional Advisory Council to Denmark, Germany, Estonia, Latvia, Lithuania, Poland, Finland and Sweden.</rec>

<rec>(4) As required by Article 3(2) of Decision 2004/585/EC, the Member States concerned determined whether the application concerning the Regional Advisory Council for the Baltic Sea was in conformity with the provisions laid down in that Decision. On 13 December 2005, the Member States concerned transmitted a recommendation on that Regional Advisory Council to the Commission.</rec>

<rec>(5) The Commission has evaluated the application by the interested parties and the recommendation in the light of Decision 2004/585/EC and the aims and principles of the Common Fisheries Policy, and considers that the Regional Advisory Council for the Baltic Sea is ready to become operational,</rec>

```
      </pr.rec>
  <pr.final>HAS DECIDED AS FOLLOWS:</pr.final>
  </pr>
  <et>
  <art>
      <art.ti>Sole Article</art.ti>
  <al>The Regional Advisory Council for the Baltic
Sea, established by Article 2(1)(a) of Decision
2004/585/EC, shall be operational as from 13 March
2006.</al>
  </et>
  <fi>
  <fi..date>Done at Brussels, 1 March
2006.</fi.date>
  <fi.sign>
  <p>For the Commission</p>
  <p>Joe BORG</p>
  <p>Member of the Commission</p>
      <fi./sign>
  </fi>
  </act>
```

Figure 2: XML annotation for European document.

III. XML RETRIEVAL MODELS

For information retrieval from document-centric XML data, the research community has exhibited a large interest in XML retrieval models. In information retrieval a representation is made from each document, which at query time is matched with the representation of the query [3] . A retrieval model (e.g., vector space model, probabilistic language model) is defined by the query representation, the document representation and the function that is used to match a document and a query [3]. While the latter retrieval model relies on a deterministic matching of query data and object data in the database, the former incorporates an element of uncertainty, i.e., documents can be retrieved even if their content representation does not exactly match the one of the query. When retrieving data from a database and one of the query conditions is not fulfilled by a data object, the object will not be retrieved. Typical query languages such as Xpath and Xquery for retrieving information from XML documents are inspired by the SQL (Structured Query Language) language and exploit Boolean retrieval, i.e., a deterministic matching of query terms and markup information. Such an approach does not allow the ranking of documents according to the relevance to the query. Typical for an information retrieval model is the relevance ranking of the retrieval results that is the consequence of a non-deterministic or probabilistic matching [3].

IV. XML RETRIEVAL MODELS FOR LEGISLATION

The legal information retrieval is an important information technology application, and it has an increasing significance. Legislative texts are currently accessible through specifically designed portal sites owned by governments or private institutions. The search engines that operate on the legal documents usually offer a full-text search (i.e., every word of the text including some metadata is indexed and can be searched). A full-text search is popular because it provides a flexible information access: the user can build any search query. When information is retrieved by using a full text search, the resulting answers of a search are ranked according to relevance to the query. The current search engines that operate on legislation allow for an extra selection of the content through filling out specific fields that represent specific structured content of the document (e.g., document title, number of an article,etc.) [3] . There is a recent trend in information retrieval to take into account the structured information of documents (e.g., as marked by XML) and especially the hierarchical logical document structure when generating the answer to a query and when computing the relevance ranking. This has several advantages. The use of the document structure allows generating a more precise answer to an information query. Instead of returning the complete document as the answer, a structural element or several elements are given. Such an approach meets the current need of users of legal information systems, who demand more precise answers to information queries . Moreover, research has only recently started to exploit the relationships between structured elements in ranking functions [1].

V. A VECTOR SPACE MODEL

In a traditionally IR system[1], queries and documents are syntactically analyzed and reduced into term (noun) vectors. A term is usually defined as a stemmed non stop-word. Very infrequent or very frequent terms are eliminated. Each term in this vector is represented by its weight. The weight of a term is computed as a function of its frequency of occurrence in the document collection and can be defined in many different ways. The term frequency - inverse document frequency (tf·idf) model is used for computing the weight [9]. Typically, the weight q_i of a term i in a document is computed as :

$$qi = tfi \cdot idfi = tfi .\log(N/n) \qquad (1)$$

where : tfi is the frequency of term i in the document (number of word occurrences in a document); this count is usually normalized to prevent a bias towards longer documents, idfi is the inverse frequency of i in the whole document collection , N is the number of all documents, n is the document frecquency (number of documents containing the word). Traditionally, the similarity between two documents (e.g., a query q and a document d) is computed according to the Vector Space Model (VSM) [8] as the cosine of the inner product between their document vectors .

$$\mathrm{sim}(d,q) = \frac{\sum_i q_i d_i}{\sqrt{\sum_i q_i^2 \sum_i d_i^2}} \qquad (2)$$

where q_i and d_i are the weights in the two vector representations. Given a query, all documents are ranked according to their similarity with the query.

VI. A VECTOR SPACE MODEL FOR XML RETRIEVAL

For this model , we first take each text node (which in our setup is always a leaf) and break it into multiple nodes, one for each word. Next we define the dimensions of the vector space to be lexicalized subtrees of documents – subtrees that contain at least one vocabulary term [5].

We can now represent queries and documents as vectors in this space of lexicalized subtrees and compute matches between them. This means that we can use the vector space formalism from (2) for XML retrieval. The main difference is that the dimensions of vector space in unstructured retrieval are vocabulary terms whereas they are lexicalized subtrees in XML retrieval [5]. In the regular vector space model, documents and queries are indexed in a similar manner, so as to produce vectors in a space whose dimensions represent each a distinct indexing unit t_i. The coordinate of a given document d on dimension t_i, , is noted $w_d(t_i)$ and stands for the "weight" of t_i in document d within a given collection. It is typically computed using a score of the tf x idf family that takes into account both document and collection statistics. The relevance of the document d to the query q, noted below *r* (q,d) , is then usually evaluated by using a measure of similarity between vectors such as the cosine measure, where:

$$\rho(q,d) = \frac{\sum_{t_i \in q} w_q(t_i) \cdot w_d(t_i)}{||q|| \cdot ||d||} \qquad (3)$$

where ti is the indexing unit, wd(t_i) is the weight of t_i in d (document), wq(t_i) is the weight of t_i in q (query). This model use as indexing units not single terms but pairs of the form (t_i,c_i), where terms are qualified by the context in which they appear. The context of a leaf node is the path from the root element to the leaf (eg.,Fig. 2 „act/pr/pr.init"). In order to identify this context of appearance, we borrow from the XPath model of XML documents – where each document is represented by a tree of nodes – its use of a path notation for navigating through the hierarchical structure of the document [6]. A structural term is a term in context (eg., „act/pr/pr.init/ THE COMMISSION OF THE EUROPEAN COMMUNITIES "). In Figure 2, the first occurrence of "fisheries" will be associated with the path "/act /ti/p " and with de path „act/pr/pr.cit". To distinguish the context in which a term is used, define a vector space in which each distinct structural term is a separate dimension . For example, in our case we have the following documents: dimension 1: "act /ti/p/fisheries", dimension 2:

[1] **Information retrieval (IR)** is a branch of searching science for: information in documents, documents themselves, metadata which describe documents, or searching within database(relational or hyper textually)

„act/pr/pr.cit/fisheries”. We suggest then changing the similarity measure accordingly. Thus, in (3) the weight of individual terms should be replaced by a weight in context that we note Wd(t_i ,c_i) [6]. This model suggest to increase the relevance score not only when a same (t_i ,ci) is found in the query and the document, but also when a same t_i appears in different but somehow related contexts c_i and c j [6]. We use cr (context resemblance), the measure of resemblance between contexts, we propose to use as measure of similarity between XML fragments and XML documents:

$$\rho(q,d) = \frac{\sum_{(t_i,c_q)\in q} \sum_{(t_i,c_d)\in q} w_q(t_i,c_q) \cdot w_d(t_i,c_d) \cdot cr(c_q,c_d)}{||q|| \cdot ||d||} \quad (4)$$

where $w_q(t_i,c_i)$ is the weights of term t_i in XML context in query q and $w_d(t_i,c_i)$ is the weights of term t_i in XML context in document d. The context resemblance function, cr, is a measure of how closely the context of a term in a query matches the context of a term in the document. One suggested measure is [5]:

$$cr(c_q, c_d) = \begin{cases} \dfrac{1+|c_q|}{1+|c_d|} & \text{if } c_q \text{ matches} \\ 0 & \text{if } c_q \text{ does not match} \end{cases} \quad (5)$$

where $|c_q|$ and $|c_d|$ are the number of nodes in the query path and document path, respectively, and c_q matches c_d iff we can transform cq into cd by inserting additional nodes. For example, if we have d0=act/ti/p/fisheries and the query q=act/ti/fisheries/,then cr(q,do)=4/5= 0.8. We impose that cr values range between 0 and 1, where 1 is achieved only for a pair of perfectly identical contexts. Thus, we see that (4) is identical to (3), in the special case of free-text where there is one unique default context. Using a vector space model for XML retrieval, I want to find the acts what contains the word „fisheries” in the tile. In this case, the query may be rewrite as : act/ti/fisheries. The document in Figure has 334 nine structural terms(documents) (Example of these d0=act/ti/p/fisheries; d1=ti/p/fisheries; d2=p/fisheries; d3=fisheries; d4=act/pr/pr.cit/cit/fisheries; d5= pr/pr.cit/cit/fisheries; d6= pr.cit/cit/fisheries; d7=cit/fisheries; d8= fisheries d9=act/pr/pr.rec/rec/fisheries d10=pr/pr.rec/rec/fisheries; d11= pr.rec/rec/fisheries; d12= rec/fisheries; d13= fisheries....). Using formulae (4) and (5), we calculate the similarities between documents d0,d1,..d12 respectiv and the original query. In this example, the highest ranking document is d0 with a similarity of 0.96.

VII. CONCLUSION

Since the last decade, XML has gained growing importance as a major means for information management, and has become inevitable for complex data representation. Due to an unprecedented increasing use of the XML standard, developing efficient techniques for comparing XML-based documents becomes crucial in information retrieval (IR) research . In legal XML retrieval not only documents but also fragments of documents are retrievable units. Therefore, most researchers are treating XML elements as independent XML documents. We can utilize this issue for another approach of term weighting, which is a crucial und still unsolved problem in semi-structured document retrieval. The Vector Space Model is one of the most popular models used in IR. It is based on the comparison of the query term vector with the document term vectors. Each term has a certain weight which reflects its descriptiveness with respect to the query or document. Extending this model to terms vectors consisting of structural terms. That is, each component of the query and document vectors contains the weight of a structural term. This extending model suggest to increase the relevance score not only when a same (ti ,ci) is found in the query and the document, but also when a same ti appears in different but somehow related contexts ci and c j. For determining similarity they suggest a context resemblance similarity measure that uses weights for both the context and the term similarity.

REFERENCE

[1] Marie-Francine Moens, „Retrieval of Legal Documents: Combining Structured and Unstructured Information”, Proceedings ELPUB2005 Conference on Electronic Publishing Kath. Univ. Leuven June 2005

[2] Joint Practical Guide for the drafting of Community legislation. Avaible: http://reterei.eu/rete/GPC_en.pdf

[3] Marie-Francine Moens, ‘XML Retrieval Models for Legislation’ in T. Gordon (ed.), Legal Knowledge and Information Systems. Jurix 2004: The Seventeenth Annual Conference. Amsterdam : IOS Press, 2004, pp. 1-10

[4] Farah, Fady, Rousselot François, “DARES: Documents annotation and recombining system—Application to the European law”, Artificial Intelligence and Law, Volume 15, Number 2, June 2007 , pp. 83-102(20), Springer 2007

[5] Christopher D. Manning, Prabhakar Raghavan and Hinrich Schütze, „Introduction to Information Retrieval” , Cambridge University Press (to appear in 2008) , Avaible : http://nlp.stanford.edu/IR-book/

[6] Carmel, D., N. Efrati, G.M. Landau, Y.S. Maarek and Y. Mass, “An Extension of the Vector Space Model for Querying XML Documents via XML Fragments”, XML and Information Retrieval (Workshop),14-25, Tampere, Finland, August 2002.

[7] Joe TEKLI, Richard CHBEIR, Kokou YÉTONGNON, "Semantic and Structure Based XMLSimilarity-AnIntegratedApproach”, 13th International Conference on Management of Data (COMAD'06), Delhi, India, December 2006.

[8] Vector space model, Wikipedia, Avaible: http://en.wikipedia.org/wiki/Vector_space_model

[9] Giannis Varelas,Epimenidis Voutsakis, Paraskevi Raftopoulou, Euripides G.M.Petrakis, Evangelos E. Milios, “Semantic Similarity Methods in WordNet and their Application to Information Retrivial on the Web”, WIDM’05, November 5, 2005, Bremen, Germany.

[10] Basic structure of legislative acts. Avaible http://publications.europa.eu/code/en/en-130300.htm