Lattice Cube Semantic Index Based Mining on XML Documents

Dr. A.M. Natarajan
Principal
Kongu Engineering College
Erode, Tamilnadu
India

principal@kongu.ac.in

K. Premalatha
Assistant. Professor / CSE-PG
Kongu Engineering College
Erode Tamilnadu
India

kpl_barath@yahoo.co.in

A. Kogilavani
Final M.E.
Kongu Engineering College
Erode, Tamilnadu
India

vani_sowbar@yahoo.co.in

Abstract - XML (eXtensible Markup Language) is fast becoming the de facto standard for information exchange over the Internet. As more and more sensitive information gets stored in the form of XML, sophisticated indexing schemes are required to speedup document storage and retrieval. XML documents can be hierarchically represented by elements. This paper describes a Lattice-map semantic indexing technique to cluster XML documents. To improve performance of information retrieval, documents can be indexed using Lattice-map technique. Similarity and Popularity operations are available in Lattice-map indexing technique and clustering algorithm is used for mining XML documents.

Index Terms : Lattice-map indexing, Lattice Cube, XML, Path-Document Matrix

1. INTRODUCTION

The increasing relevance of the Web as a mean for sharing information around the world has posed several new interesting issues to the computer science community. The traditional approaches to information handling are ineffective because they are mainly devoted to the management of highly structured information, like relational databases, whereas Web data are semi structured and encoded using different formats like HTML, XML, email messages and so on.

The eXtensible Markup Language (XML) is a W3C standard for presenting and exchanging information on the Internet. In recent years, more and more areas are adopting the XML standard to represent their information from XML documents, a number of XML query languages have been developed. Sophisticated indexing schemes have been proposed to speed up document storage and retrieval. However, because the size of XML documents is very large and the types vary typical information retrieval techniques such as LSI (Latent Semantic Index) are not satisfactory. Information retrieval on the Web is not satisfactory due to partly poor usage of structure and content information available in XML documents.

Many Web applications that process XML documents, such as grouping similar XML documents and searching for XML documents that match a sample XML document, will require techniques for clustering and classifying XML documents. It has been well-established in such fields as database management and information retrieval that the more semantics about data are understood by a system, the more precise queries can become. It is intuitively obvious that if some of the rich semantics of XML can be taken into account, it is more powerful basis of supporting the clustering and classification of XML documents for a wide variety of XML applications.

Consider a document databas (D). Each document (d) is represented in XML and it contains XML elements. Each element has zero or more terms bound to it. Typical indexing requires a frequency table that is a two-dimensional matrix indicating the number of occurrence of the terms used in documents. By generalizing this idea, this paper introduces Lattice Cube that consists of triplet (d, p, c). Here p is an XML path and c is a concept. In such context, we address the problem of clustering structurally similar XML documents. This problem has several applications like recognizing different sources providing the same kind of information and in the structural analysis of a web site. Grouping semi structured documents according to their structural homogeneity can help in devising indexing techniques for such documents, thus improving the construction of query plans. This paper describes a new Lattice-map indexing based technique which is referred to as “Lattice Cube”, that represents the triplet (d, p, c) where d represents document, p is an XML path and is concept and clustering technique that can cluster such documents semantically. Before going further, consider the following example.

T. Sobh (ed.), Innovations and Advanced Techniques in Computer and Information Sciences and Engineering, 261–266.

1.1 Motivating Example

```
(1) <section>
(2)  <section> XML is represented in a
                    Lattice-map indexing …
(3)   <section> it is a new standard …
      </section>
(4)   <section> An application is as shown in
(5)    <figure>
           http://www.a.b.c/clustering.algs
       </figure>
(6)    <caption> Clustering Algorithm
        </caption>
(7)   </section>
(8)  </section>
(9)  <section> Lattice-map indexing
      technique … </section>
(10)<verticalskip/>
(11)</section>
```

Figure 1: XML Document

Suppose that a query Q1 is posed to find all documents that describe "Indexing" in more than one sub-subsection. Notice that this type of queries asks for a specific document structure that is not for section, nor for subsection but for sub-subsections. Searching an entire XML database is costly because a word pattern for search is rarely used if we search against a large document database. That is, a word list for a document is sparse as compared to the list of words available in the database. Search for a sparse list of words is not efficient. To resolve this problem, this paper proposes a way of clustering XML documents semantically. In this way, searching can be restricted within only a cluster, instead of all documents in order to improve the performance.

1.2 Organization

The remainder of this paper is as follows. Section 2 describes preliminaries such as element paths in XML documents, path-document matrix, popularity of a path column, finding radius and center. Section 3 introduces preprocessing steps, Lattice Cube that represents a set of triplets (document d, XML element p, terms or contents c). Section 4 describes XML document clustering based on Lattice-map indexing.

2. PRELIMINARIES

This section defines the following technical terms.

Definition 1: ***(Element Content)*** An XML element contains (1) simple content, (2) element content, (3) empty content, and (4) reference content.

As an example, consider an XML document as shown in Figure 1. The element <section> in line (9) has a simple content. The element <section> in line (1) has element content, meaning that it contains two subsections as shown in lines (2) and (9). Of course, two content types can be mixed, e.g., the element <section> in line (2) contains a simple content in line (2) and also elements in lines (3)-(8). The element <verticalskip> contains empty content. The content <figure> has reference content that hyperlinks to a site.

Definition 2: (*ePath*) Element Path, called "ePath,"[8] is a sequence of nested elements where the most nested element is simple content element. For example, in Figure 1 section.section.section.figure is an ePath, but section itself is not an ePath due to the top element <section> does not have simple content An XML document is defined as a sequence of ePaths with associated element contents. An XML document database contains a set of XML documents. This paper proposes a Lattice-map index for an XML document database. In a document ePath Lattice-map index, a path column represents an ePath, and a row represents an XML document.

```
d1:
<e0>
 <e1> V1 </e1>
 <e1> V11 </e1>
 <e2>
 <e3> V2 V3 V5 </e3>
 <e4> V3 V8 </e4>
 <e5 />
<e5>V12</e5>
</e2>
</e0>
```

```
d2:
<e0>
  <e1> V1 </e1>
  <e1> V11 </e1>
  <e2>
  <e3> V3 V7 </e3>
   <e3> V3 V8 </e3>
   <e4> V9
    <e5/>
  <e6> V4 </e6>
  <e7> V6 </e7>
  </e4>
  </e2>
   <e8> V6 V12 </e8>
  </e0>
```