The 8th International Conference on e-Business (iNCEB2009)
October 28th-30th, 2009

CLUSTERING XML DOCUMENTS USING PATH-BASED APPROACH

Ei Ei Mon, Khin Nweni Tun

University of Computer Studies, Yangon
eieimonucsy@gmail.com , knntun@gmail.com

ABSTRACT

In this research, we emphasize a path-based approach to cluster the XML documents. It's structure combines with the methods of common XPath and K-means clustering of partitioning. We introduce methodology for partitioning based clustering XML documents on the basis of their structural similarities, which encodes the frequently occurring elements with the hierarchical information. There have variety of methodologies based on path-based and tree-based XML document clustering. The path based clustering provides good methodology for calculating similarities among documents. Data mining acts as a feature extractor in the clustering process. We devise an implication with frequent structure mining for searching similarities between huge numbers of XML documents' structures. Dimensional matrix is built by using extracted paths and k-means clustering algorithm is applied to create accurate clusters. Based on the idea of using path based clustering, which groups the documents according to their common XPaths, i.e. their frequent structures. The quality of clustering can be measured on the dissimilarity of document structures. The testing results are used with the Sigmod XML data.

Index Terms— *common XPath, K-means clustering, XML Document Clustering, Data mining, Frequent Structure Mining*

1. INTRODUCTION

The clustering of XML documents facilitates a number of application such as improve information retrieval, document classification analysis, structure summary, improve query processing, and XML search engine. XML is different from other web documents such as HTML or text because it contains the hierarchical structure and relationships between elements. In the future work, there will have many XML repositories. XML, like HTML, makes use of tags and attributes. While HTML predefines the meanings of tags and attributes, XML can define their own document formats and allows users to specify elements (tags) and attributes using their own words. The elements and their arrangement in the hierarchy describe the document structure and imply its semantic meanings. This property offers the advantage of defining the data semantics, while it also imposes problems in document manipulations. For example, when the XML structures contain the same data and even when they use the same document type descriptor (DTD), they may be annotated by different elements (irregular) or they may be incomplete (i.e., having non-identical tree structures). The order of where the element resides in the structure is important in determining the structural similarity between the XML document and existing clusters. As the heterogeneity of XML sources increases, the need for organizing XML documents according to their structural features has become challenging. The detection of structural similarities among documents can help in solving the problem of recognizing different sources providing the same kind of Information. Structural analysis of Web sites can benefit from the identification of similar XML documents, which can serve as the input for wrappers working on structurally similar Web pages.

The main contribution of this work is a methodology for grouping structurally similar XML documents. Data mining approach to XML document clustering is pursued. Thus, data mining is treated as a feature extractor for documents clustering. The concept of frequent tree pattern for determining XML similarity was introduced in Lee et al. (2001) [6] and Miyahara et al. (2001, 2002) [7] [8]. This approach makes use of data mining techniques to find the repetitive document structure for determining the similarity between documents. In Lee et al. (2001)[6], the structural similarity is defined as the number of paths that are common and similar between the hierarchical structure of an XML document using automata and determine the frequent path of a tree using an adapted sequential mining approach. In order to mine the frequent tree patterns or structural sequences, all XML-sequences are extracted and then mining of the common frequent XPaths. The Clustering for the whole structure of XML documents and all XPaths can sparse the feature vector. To solve this problem, this research combines the methods of common XPath and K-means clustering of partitioning that improve the efficiency for those XML documents with many different structures. In Ho-pong Leung et al. (2005)[2],Apriori has been developed for rule mining in large transaction databases. Many other algorithms developed are derivative and extensions of this algorithm. A major step forward in improving the

performances of these algorithms was made by a novel compact data structure, referred to as FP-tree, and the associated mining algorithm, FP-growth.
The rest of the paper is organized as follows. Section 2 presents the related works. In section 3, Generating Paths and Mining frequent Paths with Frequent Patten Growth are described. Section4 reports measuring similarity between XML documents and finally we conclude the paper.

2. RELATED WORKS

Recent studies have proposed techniques for clustering XML documents. In clustering XML documents need to consider both element and its structure. XML paths can represent both element tags and their position information and express the structure of XML. A path represents the tags from root node to terminal node. It includes the XML tags but also reflecting the structure of the XML documents. Using XML documents path feature is a good method to compute the similarity among XML documents. To reduce the number of path features, Ho-pong Leung(2005)[2] used apriori algorithm to find the frequent paths and take these frequent paths as XML document feature that called common Xpaths. By using all paths less than or equal to length L is a user-specified parameter value as feature vectors for XML documents , Jin-sha Yuan (2008)[3], is usually sparse due to the feature vector matrix. The similarity matrix made up of path feature vector is very big and sparse. The feature vector matrix is a high-dimensional matrix in which many entries are zero, its row numbers and column numbers are very big. To reduce the dimensionality of the vectors, Jianghui Liu et al (2004)[4] use principal component analysis (PCA) to identify significant dimensions and condense the matrix. Ho-pong Leung et al(2005)[2] use aprior algorithm to mining the frequent Xpath as feature. both use an approximate method to reduce the dimensional space.

3. GENERATING PATHS AND MINING FREQUENT PATHS WITH FREQUENT PATTEN GROWTH

The XML's hierarchical structure can be represented by a rooted labeled tree (W3C's DOM)[9]. Figure 2 shows the correspondence between the XML document and its XML tree, which is a rooted tree. Each node represents an element in the XML document and the children of each node are the subelements of that node. The tree is decomposed into XPaths to represent the structured path information called node paths of the XML document. Each path contains the node properties from the root node to the leaf node. XPath provides a way to describe the structure of a source document so that can transform the document. An XPath is formally defined as an ordered sequence of tags from a root to a leaf node which include hierarchical structure.

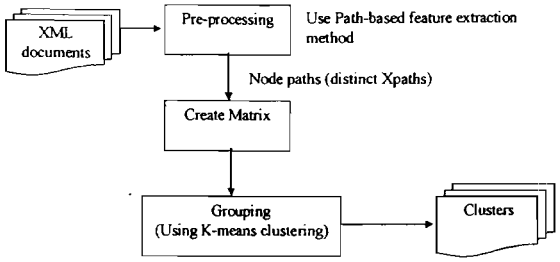

Fig: 1 Overview of the Proposed System

Duplicated XPaths in a document structure are eliminated. After the pre-processing of XML documents, documents are represented as a collection of distinct XPaths. The structural similarity between XML documents can be computed by determining the number of paths and their level of hierarchy that are similar. Similar documents can be grouped by the same cluster. For feature extraction ,use path-based feature extraction. Frequent structure mining is an implication of mining methodology for frequent structures without candidate generation. It constructs FP-tree that is used highly compact data structure with FP-Growth algorithm (Frequent Pattern Growth) to compress the original documents structures. The resulting is greater efficiency than Apriori based algorithms.

```
<SigmodRecord>
<issues>
<issue>
<volume>15</volume>
<number>2</number>
<articles>
<article>
<title articleCode="152033"> </title>
<authors>
<author AuthorPosition="03">F Andersen</author>
<author AuthorPosition="04">H Blanken</author>
<author AuthorPosition="02">K Kuespert</author>
<author AuthorPosition="01">P Dadam</author>
</authors>
</article>
</articles>
</issue>
</issues>
</SigmodRecord>
```

It is an implication of mining methodology for frequent structures without candidate generation. As shown in J. Han, J. Pei et al (2000)[5], the main bottleneck of the Aprioi-like methods is at the candidate set generation and test. This problem was dealt with by introducing a novel, compact data structure, called frequent pattern tree, or FP-

tree then based on this structure an FP-tree-based pattern fragment growth method was developed, FP-growth. FP-tree is an efficient algorithm for finding frequent patterns in transaction database and a compact tree structure is used.

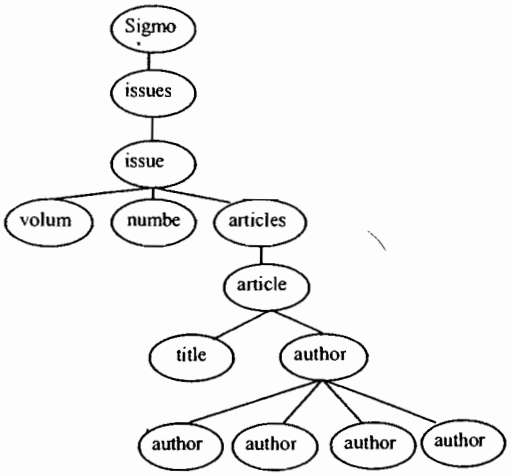

Fig: 2 Example of XML tree generation

Mining based on the tree structure is significantly more efficient than Apriori, J. Han et al (2000)[5]. FP-Growth algorithm is used highly compact data structure (Frequent Pattern Growth) to compress the original documents structures. The resulting is greater efficiency than Apriori based algorithms.

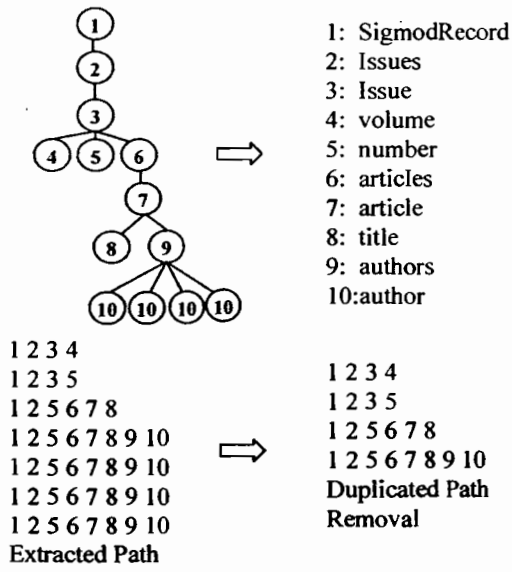

Fig: 3 Finding Duplicated Path with Frequent Pattern Growth

4. MEASURING SIMILARITY BETWEEN XML DOCUMENTS

Our clustering algorithm is based on the mining results of the CXPs. As described in the previous section, the FP mining algorithm find the maximal common paths(i.e. maximal frequent sequences) from the XML paths of the documents for comparison. If the two documents are very similar to each other, more common paths are mined and have good matches to the extracted path from the original documents. Based on the mining results, the similarity between two documents can be measured by the maximal common paths.

```
<book>
<authors>
<author>martin louis</author>
<other_author>karypis</other_author>
</authors>
<title>data mining</title>
</book>
```

(a) XML document doc1

```
<book>
    <author_list>
        <author>martin lousi</author>
        <other_author>Jason</other_author>
    </author_list>
    <title>web mining</title>
</book>
```

(b) XML document doc2

```
<book>
<author>martin louis</author>
<other_author> zhao</other_author>
<title>machine learning</title>
</book>
```

(c) XML document doc3

Figure 4: Example of three XML documents

4.1 CREATE A VECTOR MATRIX PHASE

We extract all distinct paths extracted from all XML trees with duplicate paths removed, then the dimensional feature vector matrix is defined. Each document id represented by an n-dimensional vector matrix. The feature vector matrix is defined if an XML document contains the common path is set to 1; otherwise it is set to 0. Based on the set of common path vectors collected, we compute the structure similarity between the XML documents. Feature extracted paths from three XML documents in Figure 4 is

P1=book/authors/author
P2=book/authors/other_author
P3=book/title
P4=book/author_list/author
P5=book/author_list/other_author
P6=book/author
P7=book/other_author

In Figure 4, the feature vector matrix D is :

$$D = \begin{array}{c} \\ \text{doc 1} \\ \text{doc 2} \\ \text{doc 3} \end{array} \begin{array}{c} \begin{array}{ccccccc} p1 & p2 & p3 & p4 & p5 & p6 & p7 \end{array} \\ \begin{bmatrix} 1 & 1 & 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 1 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 & 1 & 1 \end{bmatrix} \end{array}$$

4.2 Similarity Computation and Clustering Phase between XML documents

Clustering algorithms are used to find groups of "similar" data points among the input patterns. K-means algorithm is used to compute the similarities of XML documents. This is an effective algorithm to extract a given number of clusters of patterns from a training set. Once done, the cluster locations can be used to classify data into distinct classes. Partition-based clustering algorithms partition n data objects into k partitions which optimize some objective function, e.g., K-mean. Partitioned clustering algorithms are more suitable for clustering large datasets. The similarity between two XML documents is defined as using the Euclidean distance. For each pattern X, associate X with the cluster Y closest to X using the Euclidean distance:

$$\text{Dist}(X,Y) = \sqrt{\sum_{i=1}^{m} (X_i - Y_i)^2}$$

Improved XML clustering using K-means algorithm

Algorithm: XMLClustering

Input : A set of XML documents D = {d1, ..., dn},minimum support $0 \leq \delta \geq 1$

Output : clusters

(1) Prerequisite:

(a) Represent the XML documents by a set of simple XPath expressions

(b) Decide on the minimum support δ for Common XPath δ mining

(2) Discover Common XPath (set of simple XPath expressions, δ)

(a) Build the FP- Tree for the XML Documents Paths

(b) Transform each path into another representation using a mapping table

(c) Find the set of all frequent XPath expressions

(d) Find the maximal XPaths (i.e. CXPs) among the set of frequent XPath Expressions.

(3) Create a vector matrix

The feature vector matrix is defined as the following:

(1)Column is made up of XML path feature; rowis made up of XML documents.

(2) If the ith document includes the jth path, then R (i, j)=1, otherwise, R (i,j)=0.

(4) Clustering with K-means algorithm

Arbitrarily choose k objects as the initial clusters. Calculate the similarity matrix using formula (1)

Repeat

- (re)assign each object to the cluster to which the object is most similar, based on the mean value of the objects in the cluster;
- Update the cluster means; i.e., calculate the mean value of the objects for each cluster.

Until no change

5. CONCLUSION

XML has become increasingly popular and people will have strong needs for a tool to efficient and automatically retrieve the target XML documents. In order to cluster high-volume XML documents efficiently , we have proposed a matrix based on common paths information. To improved the clustering quality, an improved K-means clustering is applied in our case. Our proposed system is more suitable for the similarities between documents of different DTDs (heterogeneous XML documents) and efficient for large scale of XML documents. We believe that the proposed methods efficient and valuable for various XML based applications.

6.REFERENCES

[1] J. Han, J. Pei, Y. Yin. "Mining Frequent Patterns without Candidate Generation ". Proc. of ACM-SIGMOD, 2000.

[2] Ho-pong Leung, Fu-lai Chung, Chan, S.C.F., Luk, R. "XML Document Clustering Using Common Xpath". Proc. of the International Workshop on Challenges in Web Information Retrieval and Integration. 2005, pp.91-96.

[3] Jin-sha Yuan, Xin-ye Li, Li-na Ma. "An Improved XML Document Clustering Using Path Feature". Fifth International

Conference on Fuzzy Systems and Knowledge Discovery 978-0-7695-3305-6/08 $25.00 © 2008 IEEE DOI 10.1109/FSKD.2008.66

[4] Jianghui Liu, Jason T. L. Wang, Wynne Hsu, Katherine G. Herbert. "XML Clustering by Principal Component Analysis". Proc. of the 16th IEEE International Conference on Tools with Artificial Intelligence. 2004,pp. 658 – 662.

[5] Jiawei Han, Jian Pei, Yiwen Yin: "Mining Frequent Patterns without Candidate Generation". In Proceedings of the 2000 ACM SIGMOD international Conference on Management of Data (Dallas, Texas, United States, May 15-18,2000). SIGMOD '00. ACM Press, New York, NY, 1-12.

[6] Lee JW, Lee K, Kim W. "Preparations for semantics-based XML mining". In: Proceedings of the 2001 IEEE international conference on data mining, San Jose, CA, December, 2001, pp 345–352

[7] Miyahara T, Shoudai T, Uchida T, Takahashi K, Ueda H ."Discovery of frequent tree structured patterns in semi-structured Web documents". In: Proceedings of the Fifth Pacific-Asia conference on knowledge discovery and data mining (PAKDD), Hong Kong, China, April, 2001, pp 47–52

[8] Miyahara T, Suzuki Y, Shoudai T, Uchida T, Takahashi K, Ueda H . "Discovery of frequent tag tree patterns in semistructured Web Documents". In: Proceedings of the sixth Pacific-Asia conference on knowledge discovery and data mining (PAKDD), Taipei, Taiwan, May, 2002, pp 341–355

[9] W3C's Document Object Model (DOM) home page [http://www.w3.org/DOM/]

[10] W3C's Extensible Markup Language (XML) home page [http://www.w3.org/XML/]