CONTEXT BASED INTER STRUCTURAL MINING ON SEMI STRUCTURED DATA USING VECTOR SPACE MODEL

Antony Seba. P,
Department of Information Technology, Assistant Professor, Kamaraj College of Engineering and Technology, Virudhunagar, 626 001,India.

Dr. K. Thanalakshmi,
Department of Mathematics, Associate Professor, Kamaraj College of Engineering and Technology, Virudhunagar, 626 001,India.

ABSTRACT

Extensible Markup Language (XML) is a self describing structure which contains tags to separate elements and has hierarchies. XML is a semi structured data because it does not conform to the formal data models.XML mining is classified as XML structural mining and XML content mining.XML structural mining is further classified as intra structure and inter structural mining. This paper aims to provide inter structural mining based on context using XML PATH (XPATH). That is the XML structure is compared with more than one XML document. The given XPATH expression contains sequence of ancestor elements and this sequence is searched in the XML documents. The sequence of ancestor elements determines more specific meaning. The context is defined in the XPATH. The Vector Space Model is used for mining the XML documents which gives the similarity of a document vector to a query vector by calculating the cosine of the angle between them.

Keywords: XML, structural mining, Vector Space model

1. Introduction

XML mining is an application of data mining. XML plays a major role in Business applications collaborations and personalization. Communication takes place through XML. And large volume of data are maintained through XML. Whenever voluminous data is present then mining is essential for information retrieval. For any information retrieval system the retrieved information should be more relevant. Xml mining may be either structural mining or content mining. Again

the structural mining is further divided into inter structural mining and intra structural structure mining. Whatever may be the structural mining, the schema of the XML document plays a major role. The context aware is most important in any XML document. Because an XML document for Book library management contains author and title. Similarly CD library also contains title and author. But the context is different. Therefore context is important. For XML the context is set through the XPATH. To discover the useful information certain mathematic models are available. One among is vector space model , in which all the documents are treated as vectors.

2. **Related work**

XML plays a major role in business applications collaborations. XML mining is to discover useful information from pool of data through some process. Buhwan Jeong et. al , proposed [2] a Kernel based XML mining using vector space models and many other models.They have given a query and found the most similar , but not identical. But mining is done only on the content.

Athena Vakali [1] explains the parent child relationship in storing the XML documents. Therefore navigating through the XML document is quite easier with ancestor – descendant relationships. Also a binary model storage is explained which describes only the presence of pattern or content in an XML document

In this paper Laura et. al [3]explains the Structural mining which is concerned with determining the similar documents based on their content on static and dynamic XML documents. Association rule is applied for static XML mining. For matching semantic tags , an ontology is created in [5].Therefore automatic schema matching is done. XPATH is used for maintaining the hierarchical information [4] in the original XML document. The Information Retrieval based methods treat each document as a bag of words.

3. **Proposed Work**

This paper aims to provide inter structural mining based on context using XML PATH (XPATH). The given XPATH expression contains sequence of ancestor elements and this sequence is searched in the XML documents. The sequence of ancestor elements determines more specific meaning. The frequency of each node along with their ancestor node in XPATH is

checked in the XML documents. The sequence of the nodes in the XPATH is strictly maintained for its context. The node frequency is calculated as

$$t_{f,d} = \sum_{x \in d} f_t(x) \qquad \text{where} \quad f_t(x) = \begin{cases} 1 & if \quad x = t \\ 0 & otherwise \end{cases}$$

The document frequency is also computed. The inverse document frequency of a term t is calculated as $idf_t = \log \frac{D}{dfi}$ where D is the number of XML documents. Non frequent XML path have high idf. The weight of a term is computed using both tf and idf.

$$Weights(t,d) = tf_{t,d} * idf_t$$

tf * idf normalization is high when t occurs many times in a small set of documents. It is low when t occurs fewer times in a document or when it occurs in many documents. It is very low when t occurs in almost every document. The vector space model gives the similarity of a document vector to a query vector by calculating the cosine of the angle between them.

$$Sim(d,q) = \text{cosine } \theta$$

$$\cos\theta = \frac{d.q}{|d| * |q|} = \frac{\sum_j W_{i,j} W_{q,j}}{\sqrt{\sum W_{i.j}^2} \sqrt{W_{q,j}^2}}$$

The numerator is a dot product of two vectors, such as $\sum_{i=1ttom}(xi * yi)$ and the denominator is the product of the Euclidean length of the vector such as $|p| = \sum_{i=1}^{m} x_i^2$. sim(d,q)=1 when d=q and sim (d,q)=0 when d and q share no terms.

Q is the XPATH which is \book\author

d1,d2 are XML documents

D is the number of XML documents

XPATH	tfi	dfi	D/dfi	IDfi	Weights=tfi*IDfi

	Q	d1	d2				Q	d1	d2
\book	1	3	4	2	1	0	0	0	0
\book\author	1	0	4	1	2	0.30102	0.30102	0	1.2041

For similarity analysis, calculate the vector lengths.

$|d| = (w_{i,j})^2$

|d1| = 0

|d2| = 1.2041

|Q|=0.30102

The dot products are calculated as follows

Q.d1=0

Q.d2=0.362463

Cosine(d1)=0

Cosine(d2)=1.000003

Therefore the document d2 is similar to the XPATH.

4. **Conclusions and future work**.

In this paper , We have applied the Vector space model for context based inter structural mining on XML and the normalization is high when the XPATH occurs many times. Further We extend this work for XML documents with structural heterogeneity.

References

[1] Athena Vakali ,Barbara Catania and Anna Maddalena ,XML Data Stores:Emerging Practices Published by the IEEE Computer Society 1089-7801/05/$20.00 © 2005 IEEE ,IEEE INTERNET COMPUTING

[2] Buhwan Jeong, Daewon Lee, Jaewook Lee, and Hyunbo Cho,Towards XML Mining: The Role of Kernel ,Pohang University of Science and Technology (POSTECH)

[3] Laura Irina Rusu, XML data mining, Part 1: Survey several approaches to XML data mining, © Copyright IBM Corporation 2011, 2012

[4] Richi Nayak, The Process and Applications of XML Data Mining

[5] Yasin Ozan Kılıç, Mehmet N. Aydin , AUTOMATIC XML SCHEMA MATCHING, European and Mediterranean Conference on Information Systems 2009 (EMCIS2009) July 13-14 2009