EXTRACTING INTERESTING KNOWLEDGE FROM VERSIONS OF DYNAMIC XML DOCUMENTS

V.R.Sonawane[1], Shaila Tambe[2]

[1] Assistant Professor, [2]Student of M.E, Information Technology, AVCOE, Maharashtra, India,
vijaysonawane11@gmail.com, Shaila.waman@gmail.com

Abstract

XML has became very popular for representing semi structured data and a standard for data exchange over the web these days. The data exchanged as XML is growing continuously, so the necessity to not only store these large volumes of XML data for future use, but to mine them to discover interesting information has became obvious. The extracted knowledge can be used to make predictions. Recently, a large amount of work has been done in XML data mining. Most of the existing work focuses on the static XML data mining, while XML data is dynamic in real applications. So research has been focused more on XML documents versions and extracting information from versions of XML documents. Approach proposed in this paper is for mining association rules from changes of versions of dynamic XML documents by using information present in the consolidated delta which can be used for making future predictions.

Index Terms: *XML, Dynamic XML documents, Association Rule mining for XML*

----------***----------

1. INTRODUCTION

Recently, XML is widely used as the de facto standard for data storage and data exchange between different applications over the Web. Now a day's data storage and representation in XML format is increased which causes increasing research efforts in XML Warehousing and XML Mining. So with storing the data in a different way for easier exchange of data between different applications, extracting interesting knowledge out of the entire volume of XML data stored becomes important. The extracted knowledge might be successfully used in the decisional process to improve business outcomes.

As a result, the need for developing new languages, tools, and algorithms to effectively manage and mine collections of XML documents becomes vital. For the XML documents, finding association rules means finding relationships between simple or complex elements in the document: in other words, finding relationships between substructures of the XML document. Discovering association rules is looking for those interesting relationships between elements appearing together in the XML document, which can be used to predict future behavior of the document [6].

First we will refer to the work done for versioning XML documents, that is, methodologies that efficiently store the changing XML documents in a way that allows the fast retrieval of the historic versions. It includes solution to the issue of versioning dynamic XML documents to collect all the changes between versions in a single XML document, named consolidated delta. Finally, described proposed solution for mining association rules from changes supported by the dynamic XML documents.

The issue of versioning XML documents has been addressed by most of the methodologies which are based on the concept of the delta document (Cobena, Abiteboul, & Marian, 2005; Marian, Abiteboul, Cobena, & Mignet, 2001). Calculated and built by comparing two consecutive versions of the XML document and recording the changes that have been taking place.

2. BACKGROUND

For the warehousing and mining XML documents, at least two types of XML documents can be considered: Static XML documents, which do not change their contents and structures in time (e.g. an XML document containing the papers published in a proceedings book) and Dynamic XML documents, which change their structures or contents based on certain business processes (e.g. the content of an on-line store might change hourly, daily or weekly, depending on the customer behavior).

For Static XML documents, various methods for storing and mining them being researched and developed during the recent years. But there is still work to be done in finding efficient ways to store and mine dynamic XML documents [1].

In this paper, focus is on extracting useful information from versions of dynamic XML document using consolidated delta. The issue of versioning dynamic XML documents to collect all the changes between versions in a single XML document,

named consolidated delta. Extracted knowledge would be very useful for determining if there are any relationships between changes affecting different parts of the document and making predictions about the future behavior of the document.

Warehousing and mining dynamic XML documents can be presented in three steps as shown in Figure 1.
1. Storing multiple versions of dynamic XML documents (Fig. 1A)
2. Extracting historic changes for a certain period of time (Fig.1B) and
3. Mining the extracted changes (Fig.1C) to obtain interesting information (i.e. association rules) from them.

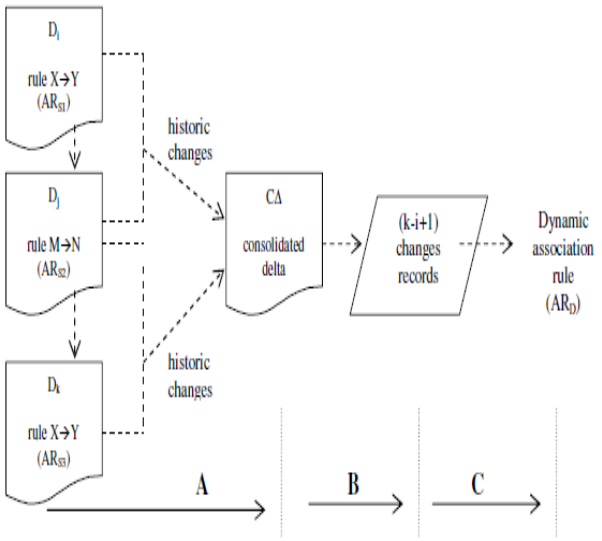

Figure1. A representation of the mining historic changes process, using consolidated delta

3. RELATED WORK

In this section, we discuss some of the existing work in the area of mining of XML documents, with the fact that there is not much work done in the area of mining changes between versions of dynamic XML documents. The existing work is more focused on determining interesting knowledge (e.g. frequently changing structures, discovering association rules or pattern-based dynamic structures) from the multiple versions of the document themselves, not from the actual changes happened in the specified interval of time.

In [2], the authors focus on extracting the FCSs (Frequently Changing Structures).

They propose an H-DOM model to represent and store the XML structural data, where the history of structural data is preserved and compressed. Based on the H-DOM model, they present two algorithms to discover the FCSs.

In [3], the authors proposed, X-Diff algorithm and it deals with unordered trees, defined as trees where only the ancestor relationship is important, but not the order of the siblings. This approach is considered to be better and more efficient for the purpose of database applications of the XML. In [3], changes in a XML document over the time are determined by calculating the minimum-cost edit script, which is a specific sequence of operations which can transform the XML tree from the initial to the final phase, with the lowest possible cost. In introduces the notion of node signature and a new type of matching between two trees, corresponding to the versions of a document, utilized to find the minimum cost matching and cost edit script, able to transform one tree into another.

Another algorithm, proposed by [4], deals with the unordered tree. But it goes further and does not distinguish between elements and attributes, both of these being get mapped to a set of labeled nodes.

In [5], the authors focus on discovering the pattern-based dynamic structures from versions of unordered XML documents. They present the definitions of dynamic metrics and pattern-based dynamic structure mining from versions of XML documents. They focus especially on two types of pattern-based dynamic structures, i.e. increasing dynamic structure and decreasing dynamic structure, which are defined with respect to dynamic metrics and used to build the pattern-based dynamic structures mining algorithm.

4. PROPOSED WORK

This paper uses the previously introduced concept of consolidated delta [1]. This is a way of storing valuable information from multi-versioned XML documents, featuring a very low degree of redundant information. The consolidated delta is built starting from the first (initial) version of the XML document, with each subsequent change which affects the document during its multi-versioned life being stored on top of the initial document. Unique identifiers are assigned to the initial nodes in the document, to be able to track the changes.

This paper proposes to build a generic algorithm for extracting association rules from the changes which affect dynamic XML documents, i.e. to discover if there are any relationships between modifications, deletions or insertions of some elements or another. The resulting rules could be very informative about how parts of the document are changing together so the user can make predictions about the future behavior of the dynamic XML document.

The algorithm for mining changes from historic versions of dynamic XML documents is an improved Apriori algorithm, modified to be applicable for XML documents; it has a preparation step and four main working steps, as mentioned below:

Preparation step:

In order to mine the actual changes which influenced the initial XML document, consolidated delta is used to extract the set of changes, using the algorithm proposed in Figure 2. The resulting document is named as ECD (the extracted changes document).

For each moment of time Ti when the document was changed, 0<i<=n, we will have a transaction containing the elements changed at the time Ti. Each combination {element – change} will actually become an item in proposed mining algorithm. During this step, also count the number of transactions in the ECD. It will be relevant for calculating the support of the association rules discovered.

For each Ti, 0<i<n
Get all nodes with timestamp Ti
For each node with timestamp Ti and delta not "unchanged"
If the delta is "modified" or "inserted"
If the node has no other children elements except stamp
Record timestamp, value and delta
End if
Else ' i.e. delta is "deleted"
Record timestamp and delta
End if
Next
Next

Fig2. Algorithm for extracting historic changes (ECD) from consolidated delta document

Step 1:

During this step, a new XML document is built, to store the number of modifications, deletions, insertions for each element Ei in ECD, together with the associated support for each pair "element-change'. The document will be named MCdoc (matrix of changes in the document), further in the paper.

The number of modifications, insertions and deletions will be stored as attributes of each element in MCdoc. Each time a new element is found in the extracted changes document, the modified, inserted or deleted attribute will take value 1. If the same element is found again to be changed during a different transaction, the corresponding attribute will be updated to reflect the current number of changes; the support of the pair {element – change} will be updated too, with regard to the total number of transactions (see preparation step).

If during the preparation step, we also include the paths of the elements in the extracted changes document, the algorithm will be able to identify only the distinct elements and their changes, so elements with same name but different positions in the hierarchy will be recorded separately.

Step 2:

The 1-large itemsets will be extracted from the document build at step 1, i.e. those items (element - change pairs) which have the support in ECD higher than the min_sup set at the beginning of the process.

Step 3:

Similar with the Apriori-based algorithm, the k-itemsets (k>1) are built starting from the 1-itemsets; for each of them the support is calculated with respect to the total number of changes from ECD. This step is repeated until all the large n-itemsets are found. This step will be influenced by the observation that any larke k-itemset (i.e. which has a support greater than minimum required) needs to have all its subsets large.

Step 4:

Based on the large n-itemsets extracted at Step 3, the association rule can be determined and their confidence can be calculated.

CONCLUSIONS

In conclusion, this paper presents a novel approach for mining changes extracted from versions of dynamic XML documents, by looking into the actual changes and into the associations between them. The information extracted would be very useful to predict the future behavior of a dynamic XML document.

REFERENCES

[1]. Rusu, L.I., Rahayu W., Taniar D., "Maintaining Versions of Dynamic XML Documents",
Proceed. of The 6th International Conference on Web Information Systems Engineering
(WISE 2005), New York, LNCS 3806, pp. 536-543, 2005
[2]. Zhao, Q., Bhowmick, S.S., Mohania, M., Kambayashi, Y., "FCS Mining: Discovering
Frequently Changing Structures from Historical Structural Deltas of Unordered XML", In
Proceedings of the 13th Conference on Information and Knowledge Management (CIKM
2004), pp. 188-197
[3]. Wang Y., DeWitt D.J., Cai J.Y., "X-Diff: An Effective Change Detection Algorithms for
XML Documents", In Proceedings of ICDE 2003, pp.519-530, IEEE Computer Society,
2003
[4]. Zhao, Q., Bhowmick, S.S., Mohania, M., Kambayashi, Y., "Discovering Frequently
Changing Structures from Historical Structural Deltas of Unordered XML", Proceedings
of ACM CIKM'04, pp.188-197, November 8-13, Washington, US, 2004

[5]. Zhao, Q., Bhowmick, S.S., Mandria, S., "Discovering Pattern-based Dynamic Structure from Versions of Unordered XML Documents", In Proceedings of the 6th International Conference on Data Warehousing and Knowledge Discovery (DaWak 2004), pp.77-86, Zaragoza, Spain, September 1-3, 2004

[6]. Web Data Management Practices-Athena Vakali & George Pallis,: "Mining Association Rules from XML Documents" pp-96- 117.

BIOGRAPHIES:

V.R Sonawane, Assistant Professor
AVCOE, Maharashtra

Shaila Tambe, M.E. (I.T), First Year, Pune University