AENSI Journals

Australian Journal of Basic and Applied Sciences

Journal home page: www.ajbasweb.com

Extracting Knowledge from XML Documents using Tree Based Association Rules

[1]S. Chandragandhi, [2]S. Shanthi, [3] Dr.L.M. Nithya

[1]*Assistant Professor, Department of Computer Science and Engineering, Rathinam Technical Campus, Coimbatore, India.*
[2]*Associate Professor, Department of Computer Science and Engineering, Rathinam Technical Campus, Coimbatore, India.*
[3]*Associate Professor, Department of Information Technology, SNS College of Technology, Coimbatore, India.*

ARTICLE INFO

Article history:
Received 23October 2013
Received in revised form 24 December 2013
Accepted 27 December 2013
Available online 1 February 2014

Key words:
XML, Approximate Answering, XML Mining, Intensional Information

ABSTRACT

The increasing use of XML technology for data storage and data exchange between applications, mining XML documents is important topic. Massive amount of XML document is available in several real applications. It cause some difficulties to non-expert users trying to access these datasets without having sufficient knowledge on their content and structure. An approach called Tree-Based Association Rules (TARs), which provides approximate and intensional information on both the structure and the contents of Extensible Markup Language (XML) documents it can be stored in XML format as well. This mined knowledge is used to provide the intensional information on both the structure and the content of the XML document. It provides approximate answers to queries. The quality of an association rule is measured by means of support and confidence. Missing value estimation is calculated for the dataset in order to improving the accuracy

To Cite This Article: S. Chandragandhi, S. Shanthi, L.M. Nithya., Extracting Knowledge from XML Documents using Tree Based Association Rules. ***Aust. J. Basic & Appl. Sci.,*** *7(14): 11-16, 2013*

INTRODUCTION

Extensible Markup Language (XML) is a flexible hierarchical for representing huge amount of data. It's nested, self-describing structure provides a simple yet flexible means for applications to exchange data. Recently, XML is passing into virtually all areas of internet application programming, producing huge amount of data in encoded format. An unstructured documents, there is a significant portion of XML documents which have only an implicit structure, (ie)their structure has not been declared in advance, Example DTD or an XMLSchema. Extracting knowledge from XML data sources turned into a very important and necessary characteristic with the continuous growth in XML data. Data mining techniques used to extracting frequent patterns (M. Mazuran et al., 2009) and providing intensional, often approximate, information both about the content and the structure of a document. Query is used to extract association rule from XML document (S.Kotsiantis et al., 2006). Missing value imputation is calculated from missing data (P. Allison, 2001).

Goal and Contributions:

There are various algorithm are available to find frequent patterns from tree based data representation(T. Asai et al., 2002). A new method called Tree-Based Association Rules (TAR) represents intensional knowledge in native XML. The idea of mining association rules (R. Agrawal et al., 1994) from XML documents by using languages (e.g., XQuery)A proposed algorithm called Tree-Based Association Rules (TAR) as a means to represent high quality knowledge in native XML. Intuitively, a TAR represents intensional knowledge in the form $S_B \rightarrow S_H$, where SB is the body tree and S_H the head tree of the rule and S_B is a subtree of S_H. The rule $S_B \rightarrow S_H$ states that, if the tree SB appears in an XML document D, it is likely that the "tree SH also appears in D. TAR provides a valid support in several cases:

1. It is used to store implicit knowledge of the Documents (ie) provides frequent patterns about the structure of the XML.
2. TAR results fast query retrieval and approximate answers.

TAR characterized by the following key aspects:

1. It is applied to XML documents, without transforming the data into any intermediate format.
2. It is general association rules, without the need to impose what should be contained in the antecedent and consequent of the rule.

Corresponding Author: S. Chandragandhi, Assistant Professor, Department of Computer Science and Engineering, Coimbatore, India.
E-mail id:chandragandhi09@gmail.com

3. It stores association rules in XML format.
4. It translates the queries on the original data set into queries on the TARs set.

Proposed System:

The proposed XML query answering support framework is as shown in figure1. The purpose of this framework is to perform mine the rules from XML and obtain intentional knowledge. The intentional knowledge is in the form of XML. It finds interesting relationships between subtrees of XML documents. The rules generated based on supports and confidence. As shown figure1, the framework is to have data mining for XML query answering support. When XML file is given as input, XML is parsed for well validness. If the given XML document is valid, it is parsed and loaded into database which can be navigated easily. The parsed XML file is given to data mining sub system which is responsible for sub tree generation and also TAR extraction.

The generated TARs are used by Query Processor System. It takes XML query from end user and makes use of mined knowledge to answer the query quickly. In XML document, we can extract two types of TAR:

- Structure TAR, It provides information only about structure of the document.
- Instance TAR, It provides information both on the structure and data values for the XML document

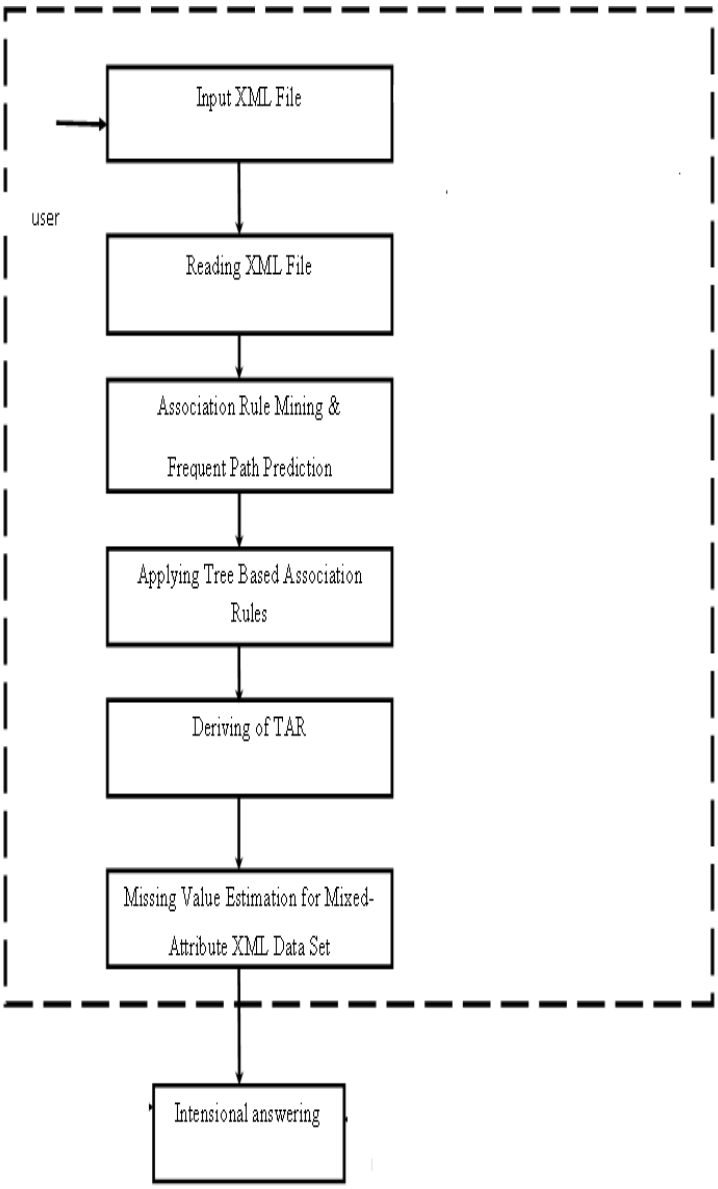

Fig. 1: Proposed XML query answering support framework

TAR Extraction:

Here Tree-based association rules are extracted from an XML document and how they are stored and used to respond to the types of queries.

Australian Journal of Basic and Applied Sciences, 7(14) December 2013, Pages: 11-16

Steps followed in TAR Extraction:

Step 1: Set Minimum Support Threshold & Confidence Threshold value
Step 2: Get the user query
Step 3: Compute Rules
Step 4: Calculate support and confidence value
Step 5: Support value>=Min_Support & Confidence >= Min_Confidence
Step 6: Mine frequent sub trees
Step 7: Extract Data

Association Rule Extraction:

It contains two steps:

1. Mining frequent subtrees from the XML document.
2. Computing interesting rules from the previously mined frequent subtrees. The problem involved in finding frequent subtrees has been discussed in (T. Asai et al., 2003) (Y. Xiao et al., 2003) (X. Yan et al., 2003)

TAR mining is different from XML association rule mining (J. Paik et al., 2005).

Tree-based association rule is characterized by two measures:

s_{Tr} support, measures the frequency of the tree S_H in the XML document.

c_{Tr} confidence, measures the reliability of a rule, that is the frequency of the tree S_H head tree, once S_B body tree has already been found.

Given function count(S, D) denoting the number of occurrences of a subtree S in the tree D and function cardinality(D) denoting the number of nodes of D, it is possible to define formally the two measures as:

$$\text{support}(S_B \rightarrow S_H) = \frac{\text{count}(S_H, D)}{\text{cardinality}(D)}$$

$$\text{confidence}(S_B \rightarrow S_H) = \frac{\text{count}(S_H, D)}{\text{count}(S_B, D)}$$

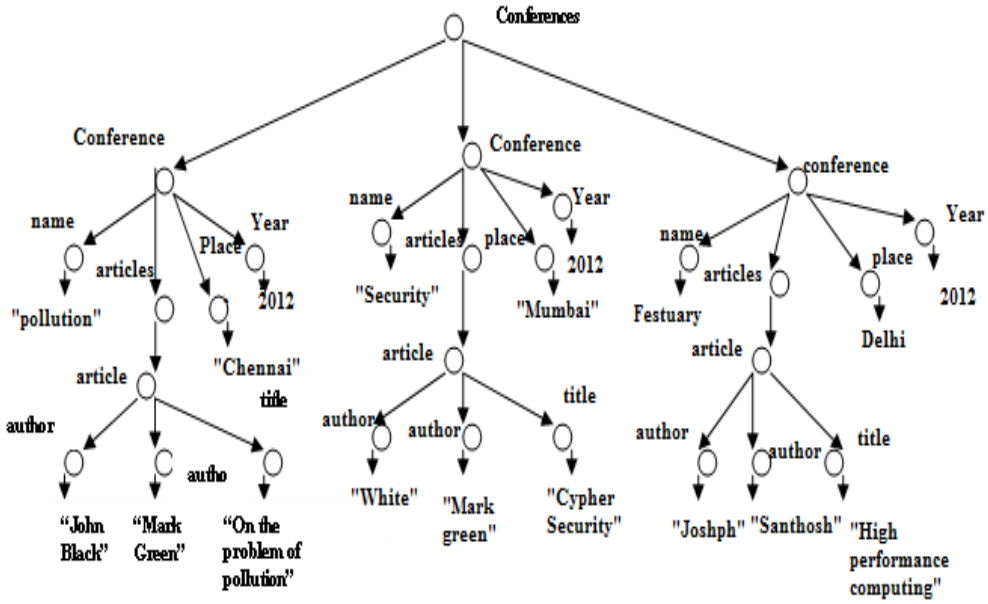

Fig. 2: Sample XML file: "Conference.xml"

Rule (1) states that if there is a node labeled with conference in the document and child node labeled year with value is \2012". Rule (2) states that if there is a path sequence of nodes: \conference \articles \ article\author, and the content of author is \Mark Green", then node authors has another child labeled author whose content is \White". Finally, rule (3), states that if there is a path contains sequence of nodes: \conference \ articles \ article and the node article has two children labeled author with contents are \Mark Green" and \John Black", then node conference probably has two other children labeled year and place whose contents are respectively \2012" and \delhi". Table 1 shows, for each one of these rules, its support and confidence.

Table 1: Support and confidence of rules in Fig 2

Rule	Rule support	Rule confidence
(1)	3/28=0.10	3/3=1.00
(2)	3/28=0.07	3/3=0.66
(3)	3/28=0.07	3/3=1.00

Experiments and Results:

The environment used to develop the prototype application includes .NET FRAMEWORK that run in Windows 7 OS. A PC with 2 GB RAM and 2.9x GHz processor is used. XQuery used for XML processing. MSSQL 2005 used as Backend for storing the details. The application GUI is as shown in fig 3.

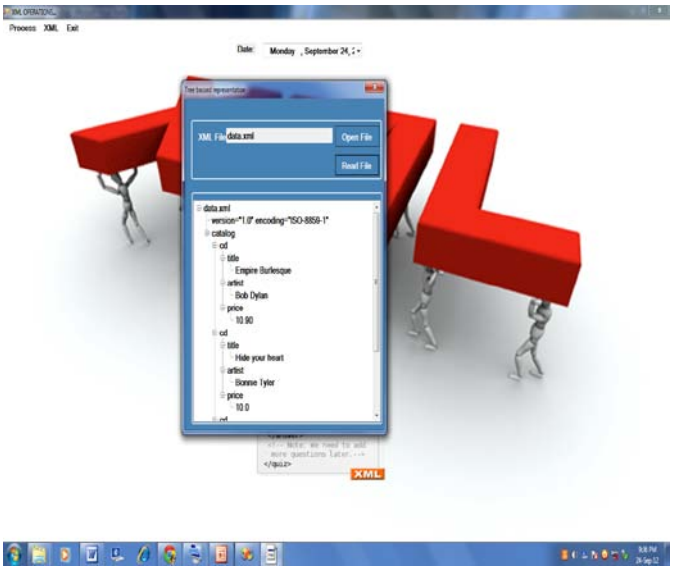

Fig. 3: XML in Tree Based Representation

Figure 3, represents XML in tree based representation. It defines (node,edge,content,tags).Finally applying Tree Based Association Rule(TAR) to the XML document. The analysis is based on the input XML file and the content of the file. The given support and confidence are considered while making the analysis. Fig. 4 shows query rules XML.A TAR is a tuple of the form Tr (S_B, S_H, sT_r, cT_r)where S_B (N_B, E_B, r_B, l_B, c_B) and S_H (N_H, E_H, r_H, l_H ,c_H) are trees and sT_r and cT_r are real numbers in the interval(0,1) representing the support and confidence of the rule.A TAR describes the co-occurrence of the two trees SB and SH in an XML document.

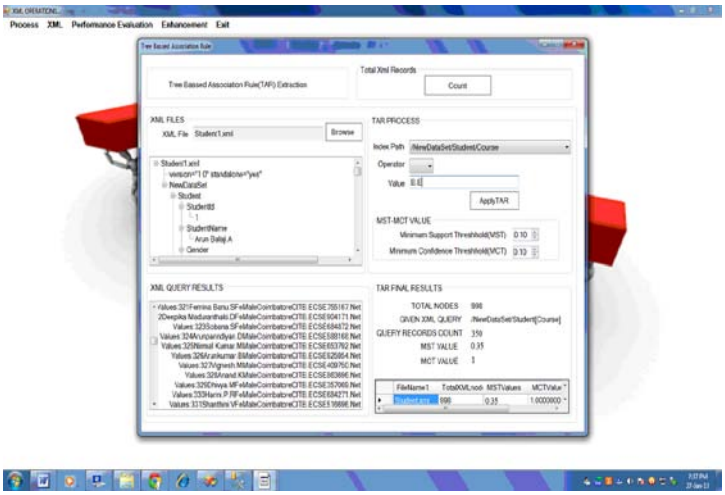

Fig. 4: Query rule for XML

As can be seen in fig. 4, the rules XML file is shown and it can be queried with various types of queries such as selection projection query, count query and Top – k query.

Final Result offers:

1)Get the gist allows intensional information extraction from an XML document, given the support, confidence and the files here the extracted TAR and their index are to be stored.

2) Get the idea allows to show the intensional information as well as the original document, to give users the possibility to compare the two kinds of information.

3) Get the answers allows to query the intensional knowledge and the original XML document. Users have to write an query; when the query belongs to the classes. It has been analyzed and it is translated to the intensional knowledge. After execution, the TAR that reflects the search criteria is shown.

Intentional answers are queried using Xquery .It provides result to the user based on three user condition.

Class 1: Retrieves the answers that satisfies AND and OR condition.

Class 2: count-queries. Used to count the number of elements having a specific content.

Class 3: top-k queries. Used to select the best k answers satisfying a counting and grouping condition.

Results:

Extraction Time:

Support defines task relevant to the query. Support determines how many frequent sub trees will be extracted. A high support threshold means for small and large sub trees, from which very less time is required to extract rules. Figure 6 defines Means of time decreases as support values increases.

Accuracy:

Missing value estimation is calculated from retrieved documents is used to evaluate the accuracy of approaches which return approximate answers. Retrieved values from XML document that contains null values that is identified and replace null value to new value. So the accuracy is improved.

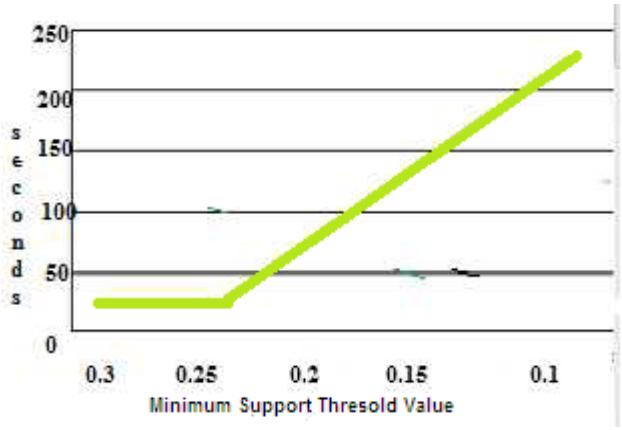

Fig. 6: Answer time

Conclusion and Future Work:

This proposed framework for extracting TAR from given XML file so as to support XML queries. Towards this end, the aim is to mine frequent association rules and store the mined content in XML format; use the TAR to support query answering or to gain information from XML databases. Future work can extend by incrementally update mined data when the original XML data sets change and how to further optimize our mining algorithm. In ongoing work, the search efficiency by handling the top-k query in the high-dimension record set.

REFERENCES

Agrawal, R and R. Srikant, 1994. "Fast Algorithms for Mining Association Rules in Large Databases," Proc. 20th Int'l Conf. Very Large Data Bases, pp: 478-499.

Allison, P., 2001. Missing Data. Sage Publication, Inc.

Asai, T., H. Arimura, T. Uno, and S. Nakano, 2003. "Discovering Frequent Substructures in Large Unordered Trees," Technical Report DOI-TR 216, Dept. of Informatics, Kyushu Univ., http:// www.i.kyushu-u.ac.jp/doitr/trcs216.pdf.

Asai,T., H. Arimura, T. Uno and S. Nakano, 2003. Discovering frequent substructures in large unordered trees.

Asai, T., K. Abe, S. Kawasoe, H. Arimura, H. Sakamoto and S. Arikawa, 2002. "Efficient Substructure Discovery from Large Semi- Structured Data," Proc. SIAM Int'l Conf. Data Mining.

Evfimievski, A., R. Srikant, R. Agrawal and J. Gehrke, 2002. "Privacy Preserving Mining of Association Rules," Proc. Eighth ACM Int'l Conf. Knowledge Discovery and Data Mining, pp: 217-228.

Gasparini, S and E. Quintarelli, 2005. "Intensional Query Answering to XQuery Expressions, " Proc. 16th Int'l Conf. Database and Expert Systems Applications, pp: 544-553.

Katsaros, D., A. Nanopoulos and Y. Manolopoulos, 2005. "Fast Mining of Frequent Tree Structures by Hashing and Indexing," Information and Software Technology, 47(2): 129-140.

Kotsiantis, S., D. Kanellopoulos, 2006. Association Rules Mining: A Recent Overview. *GESTS* International Transactions on Computer Science and Engineering.

Mazuran, M., E. Quintarelli and L. Tanca, 2009. "Mining Tree-Based Frequent Patterns from XML," Proc. Eighth Int'l Conf. Flexible Query Answering Systems, pp: 287-29.

Paik, J., H.Y. Youn and U.M. Kim, 2005. "A New Method for Mining Association Rules from a Collection of XML Documents," Proc.Int'l Conf. Computational Science and Its Applications, pp: 936-945.

World Wide Web Consortium, Extensible Markup Language (XML) 1.0,1998. http://www.w3C.org/TR/REC-xml/.

World Wide Web Consortium. XQuery 1.0, 2007. An XML query language.

Wide Web Consortium. XML Schema, 2001. http://www.w3C.org/TR/xmlschema-1/.

Xiao, Y., J.F. Yao, Z. Li, and M.H. Dunham, 2003. Efficient data mining for maximal frequent subtrees. In ICDM '03: Proceedings of the Third IEEE International Conference on Data Mining, page 379, Washington, DC, USA.

Yan, X and J. Han. Closegraph, 2003. mining closed frequent graph patterns.In KDD '03: Proceedings of the ninth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, pp: 286-295. ACM Press.