International Journal of Computer Engineering in Research Trends
Multidisciplinary, Open Access, Peer-Reviewed and fully refereed
Research Paper Volume-5, Issue-1 ,2018 Regular Edition E-ISSN: 2349-7084

Investigation of Mining Association Rules on XML Document

P.M. Gavali[1]

[1*]*Computer Science and Engineering, D.K.T.E. Society's Textile and Engineering Institute, Ichalkranji, India*

e-mail: gavalipm87@gmail.com

Available online at: http://www.ijcert.org

Received: 10/01./2018, Revised: 17/01/2018, Accepted: 18/January/2018, Published: 02/February /2018

Abstract:- XML is globally accepted format for sending the data on internet and between different applications which are running on different platforms and architectures. Due to this, the huge amount of data on the internet is in XML. Thus researchers are attracted toward XML to identify interesting findings and patterns from these documents. Many data mining algorithms have been applied to XML including clustering, classification and association rules. In this paper association, rule mining on XML document is studied. This can be used to identify what work is done in the stated field and how we can extend it further in future.

Keywords: XML, Data Mining, Association rules.

1. Introduction

Nowadays XML (eXtensible Markup Language) is widely used format for exchanging data on the internet as it is portable. Therefore most of the data available on the internet are in XML. So it became essential to process such XML documents to identify hidden and useful information from XML. In this direction, most of worked is carried out by applying various data mining algorithms. In this paper, we are concentrating on mining association rule on XML document. Rest of the paper is organized as: In point two, various techniques to identify association rule on XML used by researchers are discussed. In point three, work of various researchers is explained in detail. In point, four conclusions are given.

2. Broad Categories of Mining Association Rule on XML document

Researchers used different techniques to mine association rules from different categories. Mining association rule on XML document is broadly divided into four groups as shown in figure 1.

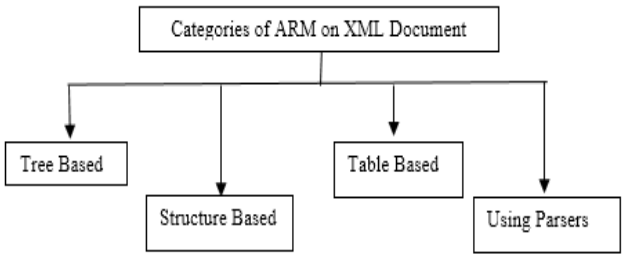

Figure 1 Broad Categories of ARM on XML Document

The tree-based approach creates a tree from XML document to simplify the identification of association rules. On this tree, basic tasks can be performed like identifying frequent itemsets and providing the abstract layer between original XML document and querying system. While structure-based method considers the structure of original XML document and their relationship with other elements in XML document to form rule. It can also be used to track the changes made in the XML document.

Some of the researchers used the variants of the table for identifying transactions from the XML document which formed the base for determining frequent items identification and identifying association rule. While parsers are used to check the quality of underlying data across DTDs and XML

© 2018, IJCERT All Rights Reserved DOI: http://dx.doi.org/10.22362/ijcert/2018/v5/i1/V5I103

Schema. It is also helpful for fast retrieval of information from the XML document.

3. Detailed Work

Win *et al*. [1] studied the association rules extraction from semi-structured XML document using Frequent Pattern-Tree based mining with divide and conquer strategy to reduce the cost evolved in candidate generation approach like Apriori algorithm. FT-Tree builds compact tree which is smaller than original database. So numbers of interactions with the database to generate candidates are minimized. Authors have also proposed FP-T growth algorithm based on XQuery assures all candidates have been identified. This method was facing problems of complex structured XML documents.

Shin *et al.* [2] reduced number of scans required for identifying association rule from the set of the XML document. And also it works well with complex structured XML document. The author proposed a Hierarchical layered structure of PairSet to reduce multiple data scans required for identifying association rule from the set of complex structure XML documents representing the same type of information. By converting tree-structured XML document into PairSet and manipulating it using Cross Filtering algorithm reduced number of iteration required for identifying association rule.

Gongxing [3] studied distributed XML documents to find frequent global patterns from the distributed environment. Proposed a FreqtTree using DOM to find frequent items in a local XML document. This is further enhanced to DFreqtTree to find frequent items in global XML document

Mahalakshmi *et al*. [4] studied XML documents without DTD and XML schema because they want to identify an efficient answering system. Authors proposed a system using Tree based association rules approach in which original XML document is given to SAX parser to identify node name and values associated with it. Such extracted nodes are considered to create LC-RS tree. Using this LC-RS tree possible frequent sub-trees are identified and stored in new XML document to provide required data to the user inefficient and more accurate manner.

Thangarasu and Sasikala[5] used Tree based association rule for identifying intentional information for both structure and contents. The result of this process is stored in XML format for further processing.

Combi *et al*. [6] studied approximate querying XML document to find structure and value rule with ordered subset which includes all possible rules. Structure rules are associated with the structure provided by the XML document. For example <phdstudent> is child with <salary>. So rule provided through the structure is phdstudent->salary. Value rule is associated with similarity of value provided in XML data concerning given query. For example research group of CV has 10000 salaries. So rule provided through XML document is ResearchGroup(CV)->10000. According to authors, this work can be extended to define support and confidence on provided rules.

Rusu *et al.* [7] studied effect done in already extracted rules due to changes made in XML document's data or structure. Authors used consolidated delta to store multi-versioned XML-document. Periodically modifications made in XML file are tracked to identify variable association rules using the algorithm for identifying variable association rule for dynamic XML document.

Khaing and Thein [8] studied mining association rule in a large amount of XML data which may be used to reduce memory storage size and to produce association rules in less response time by using Efficient Algorithm Rule Mining algorithm. Authors proposed EARM uses a binary transaction table on which basic apriori algorithm is used to identify association rules.

Li *et al*. [9] studied mining association rule because they want to identify efficient algorithm to identify association rules from XML document. Authors have proposed index table to extract the transaction and item information with node encoding technique which efficiently provides results camped to the method proposed in a tool for removing association rule from XML.

Sasikala and Premalatha [10] have proposed a methodology to modify the index table using UID for each node. This UID is considered to build index table, unlike the previous methods. The mining process is applied to the modified index table. For mining, all UID having support greater than minimum support are considered. The result showed time taken by proposed system is considerably low.

Wang and Cao [11] studied mining the XML document because they wanted to identify a solution to mine complex structured and irregular XML document which was not possible with the usage of the Apriori algorithm. Authors proposed a method in which irregular and complex XML document are converted into the simple and regular document using XSLT. This modified XML document is taken for further query processing using XQuery.

Porkodi *et al*. [12] have studied mining association rule from XML document using XQuery because they wanted to work on large XML files and wanted to improve time. Authors have proposed XML Query Association Rule Mining (XQARM) framework with XQuery and .NET based implementation consisting of three phases. In first XQuery phase, XML data files are stored in the table, and XQuery applied on the table. In second phase filtering and converting occurred items to the binary matrix is done. In the third phase, actual association rule identified which satisfies minimum support and confidence. Analyzed results indicate it worked better for large XML document within time.

© 2018, IJCERT All Rights Reserved

Bodke and Kumar [13] studied retrieving essential data from XML document to build XML query answering system. Authors used Tree based association rule for providing structure and content information necessary for query answering which provides rapid and more accurate results for various types of queries including top-k and where clause including query.

Abazeed et al. [14] studied java based FLEX algorithm to improve execution timing of it. Authors have proposed MFLEX algorithm which uses Simple API for XML and Document Object Model to provide improved execution timing.

Shahriar and Anam[15] proposed a theoretical model to prove the impact of quality of data in XML using DTD/Schema on data mining and use of data mining to measure quality data of XML using XML constrains.

Rathi *et al*. [16] studied large XML document to speed up frequent itemset mining process. Authors have proposed a methodology to handle large XML document using high performance and low-cost computing GPU which showed a significant speed up with the large dataset.

Bei *et al.* [17] have studied XML query to provide the recommendation to users for removing the unwanted rules and getting more interesting rules from users point in a rapid manner.

Iqbal *et al.* [18] have studied sensitive issues of rules to avoid sensitivity information disclose through sensitive rules in an ARM. Authors have proposed a PPDM model and bayside network to hide sensitive rules which disclose the privacy of information. Structural Transitional Itemset and Binary Itemset are used to generate association rules. By using bayside network sensitivity of data is decided and modification is done to avoid information disclose.

Suganya[19] studied XQuery for generating query-answering system using Tree based association rule. The XML document is first converted into TAR file on which query fired to provide more accurate results.

4. Conclusion and Future Scope

As XML document is globally accepted standard to transfer information between different applications and architecture, much of the data is available for mining the hidden information from such files. The various researcher used various techniques for doing the same. In this paper, we focused on association rule mining and its variants used to work on XML document. Many of this method provided the good result to find association rule from such data. Further, this work can be taken to GPU and deep learning to speed up and understand the hidden information.

5. References

1. Chit Nilar Win, Khin Haymar Saw Hla, "Mining frequent patterns from XML Data".

2. Jun Shin, Juryon Paik, and Ungaro Kim," Mining Association Rules from a Collection of XML Documents using Cross Filtering Algorithm", International Conference on Hybrid Information Technology, 0-7695-2674-8/06,2006

3. Wu Gongxing," A Study on the Mining Algorithm of Fast Association Rules for the XML Data", Proceedings of the Third International Conference on Web Information Systems Engineering (Workshops) 0-7695-1754-3/02,2002

4. S. Devi Mahalakshmi, Dr K. Vijayalakshmi, Dr K. Muneeswaran, G.Priyanka," Mining Intensional Information for answering XML-Queries using Tree-based Association Rules Approach",

5. S.Thangarasu, D.Sasikala," Extracting Knowledge from XML Document Using Tree-based Association Rules", 2014 International Conference on Intelligent Computing Applications, 978-1-4799-3966-4/14,2014

6. Carlo Combi, Barbara Oliboni, Rosalba Rossato," Querying XML documents by using association rules", Proceedings of the 16th International Workshop on Database and Expert Systems Applications (DEXA'05) 1529-4188/05,2005

7. Laura Irina Rusu, Wenny Rahayu, David Taniar," Extracting Variable Knowledge from Multiversioned XML Documents", Sixth IEEE International Conference on Data Mining - Workshops (ICDMW'06) 0-7695-2702-7/06

8. Myint Myint Khaing, Nilar Thein," An Efficient Association Rule Mining For XML Data", SICE-ICASE International Joint Conference 2006,5782-5786, Oct. 18-21, 2006 in Bexco, Busan, Korea

9. Xin-Ye Li, Jin-Sha Yuan, Ying-Hui Kong," Mining Association Rules from XML Data with Index Table", Proceedings of the Sixth International Conference on

© 2018, IJCERT All Rights Reserved

Machine Learning and Cybernetics, pg.no.3905-3910, Hong Kong, 19-22 August 2007

10. D. Sasikala, K. Premalatha," Mining association rule from XML Document using modified index table", 2013 International Conference on Computer Communication and Informatics (ICCCI -2013), Jan. 04 – 06, 2013, Coimbatore, INDIA
11. Xinwei Wang and Chunjing Cao," Mining Association Rules from Complex and Irregular XML Documents using XSLT and XQuery", International Conference on Advanced Language Processing and Web Information Technology, 978-0-7695-3273-8/08,2008
12. R.Porkodi, V.Bhuvaneswari, R.Rajesh, T.Amudha," An Improved Association Rule Mining Technique for XML Data Using XQuery and Apriori Algorithm", 2009 IEEE International Advance Computing Conference (IACC 2009),pgno.1510-1514 Patiala, India, 6-7 March 2009
13. Miss. Ujwal Arjun Bodke, Santosh Kumar," 2015 International Conference on Computing Communication Control and Automation", 978-1-4799-6892-3/15,2015
14. Ashraf Abazeed, Ali Mamat, Md Nasir Sulaiman, Hamidah Ibrahim," Scalable Approach for Mining Association Rules from Structured XML Data" 2009 2nd Conference on Data Mining and Optimization 27-28 October 2009, Selangor, Malaysia.
15. Md. Sumon Shahriar, Sarawat Anam," Quality Data for Data Mining and Data Mining for Quality Data: A Constraint-Based Approach in XML", 2008 Second International Conference on Future Generation Communication and Networking Symposia, pg.no.47-49.
16. Sheetal Rathi, C.A Dhote, Vivek Bangera," Speeding up Frequent Itemset Mining Process on XML Data using Graphic Processor", IEEE, 978-1-4799-4236-7/14
17. Yijun Bei, Gang Chen, Lihua Yu, Feng Shao, Jinxiang Dong," XML Query Recommendation Based On Association Rules", Eighth ACIS International Conference on Software Engineering, Artificial Intelligence, Networking, and Parallel/Distributed Computing, 0-7695-2909-7/07, 2007.
18. Khalid Iqbal, Sohail Asghar, Simon Fong," A PPDM Model Using Bayesian Network for Hiding Sensitive XML Association Rules", IEEE, 978-1-4577-1539-6/11, 2011.
19. I.Suganya, N.Velmurugan, Dr.P.Ganeshkumar,"XML Query-Answering Support System using Association Mining Technique".

© 2018, IJCERT All Rights Reserved