© 2016 IJSRSET | Volume 2 | Issue 6 | Print ISSN: 2395-1990 | Online ISSN : 2394-4099
Themed Section: Engineering and Technology

XML Data Analysis: Recent Review in Scope of Association Rule Generation

Vinod Prajapati, Dr. Anjana Pandey

Information Technology, University Institute of Technology RGPV, Bhopal, Madhya Pradesh, India

ABSTRACT

Revealing issues with current framework is itself a critical assignment. A review taken out for revealing issues related with Association standard mining on XML data. Preparatory essential ideas of Association rule mining is given in this work. Mining enormous amount of data, association rule mining have been demonstrated a powerful idea. Amid late years, the vast majority of the overall information exchanges are finished with XML (eXtensible Markup Language). Numerous empowering techniques have been distinguished and produced for mining XML data. In this paper, the idea of XML data examination is compressed and its importance towards association rule extraction has been represented. We have cantered a variety of strategies and methodologies of the examination, which are useful and set apart as the imperative field of XML data investigation. This work gives a study of different association rule strategies connected effectively on XML information since last one decade.

Keywords : Association Rule Mining, Semi Structures Data, XML, Graph, Tree, Frequent Pattern

I. INTRODUCTION

Data mining offers awesome guarantee in helping analyst reveal patterns hidden in their data that can be utilized to anticipate the behavior of clients, products and procedures [2, 3,7]. With the late advancement of data warehousing and OLAP innovation, organizing information at various levels of abstraction has been a typical practice [14]. Therefore, in this work, we assume such concept taxonomies exist, and our study is focused at the association rule requirement, the efficient methods for semi structured database rule mining. There are several possible challenges to explore efficient mining of XML [5, 9] data set association rules, 1) Multiple scans of transaction database, 2) Large amount of candidates, 3) Time taking support counting for candidates.

In this work, we considered the problem of Mining Association Rules between items in XML document [4, 10]. The increasing utility of XML technology for data storage and data trade between applications, the subject of mining XML reports has turned out to be more researchable and imperative theme [1,8]. The chief reason for this study is applying association principle calculations specifically to the XML records. The point of XML mining is to incorporate the rising XML innovation into information mining innovation. We presented association rules from native XML reports and examine the new difficulties and opportunities. However, efficiency and easy is still a challenge for further enhancement. However, proficiency and effortlessness is still a boundary for further advancement. Typically, pre-processing or post-processing is required for mining XML information [15], such as transforming is also considered in the work.

II. METHODS AND MATERIAL

Literature Survey

There are investigated novel and signification contributions from last one decade in the scope of XML data association rule mining.

2.1 Association Rule Mining from XML Data

Ding and Gnanasekaran [11] looked at the various approaches for association rule mining from XML data, presented an implementation of the Apriori and the FP-

Growth (Tree) algorithms using JAVA for this task, and compare their performances. They concluded that FP-Growth algorithm is normally faster than the Apriori algorithm. The presented FP-Growth algorithm receives isolate and-overcome methodology. To start with it registers the successive things and speaks to the frequent items as a packed database as a tree called frequent-pattern tree, or FPtree. The association mining is performed on this tree. This implies the dataset D should be examined just once. Additionally this calculation does not require the candidate itemset era. So it is regularly ordinarily speedier than the Apriori calculation. XQuery is intended to be a universally useful XML inquiry dialect, it is frequently extremely hard to actualize complicated calculations. So far only the Apriori algorithm has been implemented by using XQuery [12]. That is why, the presented approach for XML rule mining is to use programs written in Java to work straightforwardly with XML records. This offers more adaptability and performs all around contrasted with different strategies. For lower estimations of minimum support, it is required to have numerous frequent itemsets,and this number will decrease as the minimum support increases. So the running time decreases as the minimum support increases. The large gap between the Apriori and the FP-Growth at lower values of minimum support was caused by the large number of candidate itemsets created in Apriori.

2.2 Mining tree-based association rules from XML documents

Mazuran et al [12] described an approach to mine Tree-based association rules from XML documents. The proposed algorithm that that stretches out PathJoin and permits us to extricate frequent tree-based association rules from XML dataset. Such rules provide information on both the structure and the content of XML documents; moreover, they can be stored in XML format to be queried in future. The main goals of the work was: 1) we mine frequent association rules without imposing any a-priori restriction on the structure and the substance of the rules; 2) we store mined data in XML position; as an outcome, 3) we can viably utilize the extracted learning to pick up information, by using inquiry dialects for XML, about the main datasets where the mining calculation has been connected. They have built up a C++ model that has been utilized to test the adequacy of our proposition.

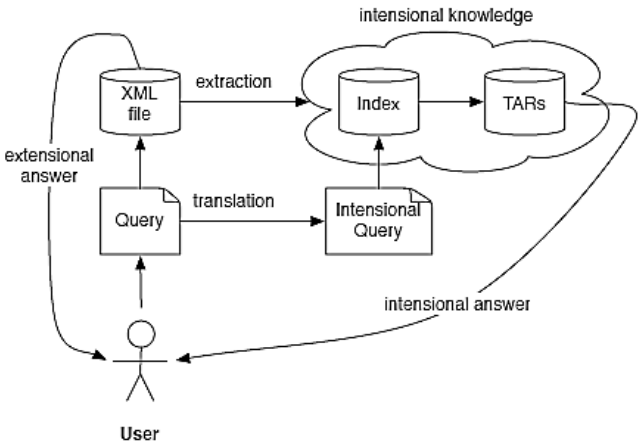

Figure 1. The architectural process of the work

2.3 Mining Association Rules from XML Data using XQuery

Wan [6] show that extracting association rules from XML documents without any preprocessing or post-processing utilizing XQuery is conceivable and analyze the XQuery usage of the understood Apriori algorithm. Moreover, they recommend highlights that should be added into XQuery keeping in mind the end goal to make the implementation of the Apriori algorithm more efficient. We see that the performance of the XQuery execution is reliant on the quantity of expansive itemsets found and the span of the dataset. We likewise see that most of the ideal opportunity for the XQuery execution is spent in numbering backing to discover expansive itemsets. The XQuery execution requires that for each itemset in the Candidate set, it will read the database once to acquire the support value. In this way, the quantity of times expected to examine the database to get the support count discovering expansive itemsets is O(2m), where m is the length of the maximal substantial itemset. This implies the entire database should be perused for every hopeful itemset. The consequences of our tests demonstrated that XQuery is more reasonable for mining information from little datasets.

2.4 Extracting Knowledge from XML Document Using Tree-based Association Rules

Thangarasu and Sasikala [13] presented a work with the goal to experimentally evaluate association rule mining method in the view of XML databases. This is the extension of work presented by Mazuran et al [12]. In this work, an approach is proposed to mine XML documents using Tree-based Association Rules. They

achieved by extracting intentional knowledge from both the structure and content of the XML document. This knowledge will improve the query answering time and easy access of the content in the XML instead of scanning entire documents. Characteristics of mining procedures are 1) input is taken as such as the XML document, 2) it provides general Association Rules without considering antecedent and consequent of the rule, and 3) it stores output in XML format. They concluded that the extraction time depends on the number of nodes in the XML data set and the number of nodes generated in the frequent subtree based on the support value. The emphasis is more on the structure rather than the content. Here to generate intentional knowledge from XML documents the frequent structure is only used. The association rules are generated for the frequent structure. Due to this, some of the interested rules may be missing. As it is generating the knowledge based on frequent structure, sometimes there is a chance that some of the tags will not be included.

2.5 An Efficient Association Rule Based Clustering of XML Documents

Muralidhar and Pattabiraman [16], association rule based mining discovers the temporal associations among XML documents. Be that as it may, this sort of data mining is not adequate to recover the properties of each XML archive. Finding the properties for set of comparative documents is preferable thought somewhat over to discover the property of a solitary document. Henceforth, the key commitment of the work is to locate the important clustered based relationship by association rule based grouping. Along these lines, this paper proposes a crossover methodology which finds the continuous XML reports by association rule mining and after that discover the cluster of XML records by established k-means calculation. All the rare items are disregarded in our philosophy since apriori has delivers just the regular archives.

III. RESULTS AND DISCUSSION

3.1 Association Rule Mining in XML databases: Performance Evaluation and Analysis

Kaur and Aggarwal [17], generating strong association rules depends on the association rule extraction method for example Apriori algorithm or FP-growth etc and the evolution of the rules by different interestingness measure such as support-confidence. The objective is to tentatively assess association rule mining approaches with regards to XML databases. Methods are actualized utilizing Java. For trial assessment distinctive XML datasets are utilized. Apriori and FP Tree calculation have been executed and their execution is assessed widely. Pre-processing and transformation of the dataset are done. During the transformation steps, Conversion of XML to CSV (Comma Separated Values) and Conversion of XML to ARFF (Attribute Relational File Format) is done. Apriori and FP Tree are applied on the different XML datasets at different levels to analyze the performance (Support, Confidence, Execution time etc.) of each algorithm at different levels. Apriori and FP-Tree generates same number of rules for each level in four different datasets. It has been predicted that FP-Tree takes less time in generating rules as compared to Apriori at each level of four datasets. When all association rule algorithms(Apriori, FP-Tree, Generalized Sequential Pattern[GSP], Predictive Apriori and Tertius) are applied on four datasets, Predictive Apriori generates maximum number of rules at minimum support and at maximum confidence as compared to all other algorithms, the sequence of generating max rules by algorithms are(Predictive Apriori > Tertius > Apriori and FP-Tree > GSP).

There is represented a comparison table (Table 1.1) which compares aforesaid six core related methods with various underlying parameters.

Table 1. Comparison of core XML based methods with various parameters

Method	Language	Data Structure	Compared with	Preprocessing Technique	Purpose	Output Style	Persistency?	Outcomes
[11]	JAVA	FP Tree	XML Xquery based java (apriori)	XQuery	Frequent item generation	Visual	No	FP Tree works well in lowest support value
[12]	C++	FP Tree		PathJoin	Answering the query	XML	Yes	FP Tree is suitable for carry XML structure

								and contents
[6]				XQuery	Frequent item generation	Visual	No	XQuery is feasible to use with apriori
[13]		FP Tree	apriory		Emphasis on xml structure rather than contents	XML	Yes	Tree structure is the best suitable for answering
[16]		apriori	Hybrid with K-means clustering		Find frequent properties of XML document structure	Visual	No	Hybrid approach improve efficiency by reducing search space
[17]		Compare: Apriori, FP-Tree, Generalized Sequential Pattern[GSP], Predictive Apriori and Tertius		CSV and ARFF	To see the performance of various algorithm with variety of parameters	Visual	No	The sequence of generating max rules by algorithms are(Predictive Apriori > Tertius > Apriori and FP-Tree > GSP).

There is vast scope of xml data mining and found some significant work where xml document and structure has been used under data mining process. Some of important contributions are:

3.2 XML Based Pre-processing and Analysis of Log Data in Adaptive E-Learning System: An Algorithmic Approach [18]

Authors have taken adaptive E-learning as a problem domain and given emphasis on the automatic skill of underlying system to calculate the candidate priority and preference. It helps to design system more interactively and arrange the learning stuff according to most promising learning style. They have suggested using web log analysis and saving the behaviors as per session wise and which is letter classified and represents in XML formats. Authors adopted Felder-Silverman Learning Style Model (FSLSM) learning model to apply on XML format data and shown better results.

3.3. A fast approach to identify trending articles in hot topics from XML based big bibliographic datasets [19]

Authors have considered the applicability of XML document in Meta data presentation and mining. In this view they have taken large bibliographic datasets while considering various underlying tags such as publishing period, caption, title, article, price, country and similar many more. Here, all the data are being presented as XML form and further many more analysis can be applied easily. There is also associated ontological data relationship with xml relation to given better results.

The main aim of the paper was to represented a framework to support large data processing and in this view, they accommodate Map reduce method and hadoop underlying system. The results shown that the performance of xml data kind of processing with map reduces give satisfactory results. The Figure 1.2 depicts the ontological matching.

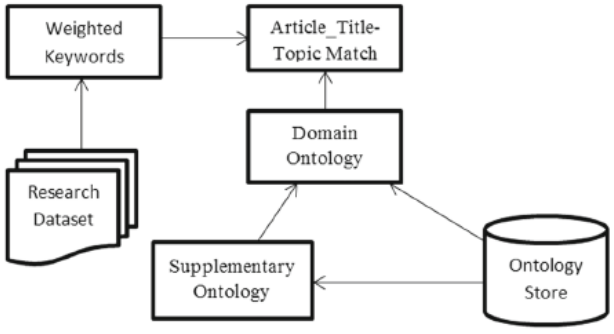

Figure 2. Ontological Matching

3.4 Clustering XML documents by patterns [20]

They presented a solid background as xml is efficient and easy technique for mining, representing, analyzing semi-structured data. In this view, authors investigated its versatile applicability in cluster analysis. Authors tried to shift pairwise comparison computation to pattern based computation through xml in clustering. They presented xml oriented pattern based document clustering. Here, xml is used to compute the maximal frequent item and its path and then combined together. The results shown that its required less parameters and less processing compare to ordinary clustering methods.

IV. CONCLUSION

Data mining has proven its utility in the area of data science since last two decades and still growing whereas association rule generation has unique applications and here, we tried to investigate its variety in various data format. Importantly, XML data representation is an agreed upon and widely used data representation and the presented survey focused around this. During the survey, we focused on variety of parameters such as underlying implementing language, pre-processing methods, technique hybridization, purpose, scope, output format etc. The conclusion is that still there is a gap for generating association rules from XML format with minimum database scan an area is open for improvement.

V. REFERENCES

[1] Vivek T. and Thakur RS, "A level wise Tree Based Approach for Ontology-Driven Association Rules Mining", CiiT International Journal of Data Mining and Knowledge Engineering, Vol 4, No 5, 2012.

[2] Agrawal R. and Srikant. R, "Fast Algorithms for Mining Association Rules in Large Databases" Proceedings of the 20th International Conference on Very Large Data Bases, pp. 478-499, 1994.

[3] Larose, Daniel T. Discovering knowledge in data: an introduction to data mining. John Wiley & Sons, 2014.

[4] S. Vurukonda, G. Ranadheer, B. Mounika, and S. Reddy, "A Survey on Tree based Association Rules (TARs) from XML Documents", Proceedings of International Journal of Research and computational Technology, vol. 5, 2013.

[5] M. Mazuran, E. Quintarelli, and L. Tanca, "Data mining for XML Query- Answering Support", IEEE Transactions on Knowledge and data Engineering, vol.24, pp. 1393-1407, 2011.

[6] W. Wan and Gillian Dobbie. "Mining association rules from XML data using XQuery", In Proc. of the 2nd Ws on Australasian information security, Data Mining and Web Intelligence, and SW Internationalization, pp. 169-174, 2004.

[7] J. Han and M. Kamber, Data Mining: Concepts and Techniques. Morgan Kaufman Publisher, 2001.

[8] Al-Maolegi, Mohammed, and Bassam Arkok. "An improved apriori algorithm for association rules." arXiv preprint arXiv:1403.3948 ,2014.

[9] Yue XU, Gavin SHAW, Yuefeng LI, "Concise Representations for Association in Multilevel Datasets,"Systems Engineering Society of China & Springer-Verlag, vol.18 (1), pp.53-70, 2009.

[10] Vivek T. & Vipin T."Association Rule Mining- A Graph based approach for mining Frequent Itemsets" IEEE International Conference on Networking and Information Technology (ICNIT 2010) , pp. 309-313, Manila.

[11] Ding Q. and Gnanasekaran S., "Association Rule Mining from XML Data." In DMIN, pp. 144-152. 2006.

[12] Mazuran, M., Quintarelli, E., & Tanca, L., "Mining tree-based association rules from XML documents." Proceedings of the Seventeenth Italian Symposium on Advanced Database Systems (SEBD), pp. 109-116. 2009.

[13] Thangarasu, S., and D. Sasikala. "Extracting Knowledge from XML Document Using Tree-Based Association Rules." Intelligent Computing Applications (ICICA), 2014 International Conference on. IEEE, 2014.

[14] V. Tiwari & Thakur RS, Contextual Snowflake Modeling for Pattern Warehouse Logical Design, Sadhana - Academy Proceedings in Engineering Science, Vol.40, Issue 1, PP. 15-33, 2015, Springer.

[15] L. Feng, T. S. Dillon, H. Weigand, and E. Chang. An xml-enabled association rule framework. In International Conference on Database and Expert Systems Applications (DEXA '12), pp. 88-97, 2012.

[16] Muralidhar, A. and Pattabiraman, V., An Efficient Association Rule Based Clustering of XML Documents, 2nd International Symposium on Big Data and Cloud Computing (ISBCC'15), Procedia Computer Science, vol. 50, pp.401-407, 2015, Elsevier.

[17] Kaur, G. and Aggarwal, N., Association Rule Mining in XML databases: Performance Evaluation and Analysis. International Journal of Computer Science and Technology (IJCST), Vol.1, Issue 2, 2010.

[18] Kolekar, S.V., Pai, R.M. and Pai, M.M., 2015, September. XML Based Pre-processing and Analysis of Log Data in Adaptive E-Learning System: An Algorithmic Approach. In International Conference on E-Learning, E-Education, and Online Training (pp. 135-143). Springer International Publishing, Springer, 2016.

[19] Swaraj, K.P. and Manjula, D., 2016. A fast approach to identify trending articles in hot topics from XML based big bibliographic datasets. Cluster Computing, Vol. 19, No. 2, pp.837-848, Springer , 2016.

[20] Piernik M, Brzezinski D, Morzy T. Clustering XML documents by patterns. Knowledge and Information Systems. Pp. 185-212, Vol. 46, Issue 1, Springer, 2016.