Available Online at www.ijcsmc.com

International Journal of Computer Science and Mobile Computing

A Monthly Journal of Computer Science and Information Technology

ISSN 2320–088X
IMPACT FACTOR: 6.017

IJCSMC, Vol. 6, Issue. 2, February 2017, pg.75 – 80

Answering XML Query Using Tree Based Association Rule

Ms. Poonam R. Maskare
Department of Computer Science & Engineering, Sant Gadge Baba Amravati University, India
poonammaskare@gmail.com

Abstract----- XML has gained popularity for information representation, exchange and retrieval. As XML material becomes more abundant, its heterogeneity and structural irregularity limit the knowledge that can be gained. The utilisation of data mining techniques becomes essential for improvement in XML document handling. In this work we describe an approach based on Tree-based Association Rules (TARs) mined rules, which provide approximate, intensional information on both the structure and the contents of XML documents, and can be stored in XML format as well. This mined knowledge is later used to provide: (i) a concise idea – the gist – of both the structure and the content of the XML document and (ii) quick, approximate answers to queries. In this work we focus on the second feature. A prototype system and experimental results demonstrate the effectiveness of the approach.
Keyword---- Extensible markup Language (XML), Association Rule mining, TAR extraction, Query Answering, Intentional knowledge

I. Introduction

The Web is an immense and dynamic collection of pages and services that includes countless hyperlinks, thus, it provides a rich and diversified data mining source. In a current technology the database research field have concentrated on xml as a flexible hierarchical model which is suitable to present large amount of data with no fixed & absolute structure[4]. Hence there is ability to extract knowledge from xml so that decision support becomes going increase & desirable. Keyword based Search & Query answering are the reasons to access XML documents. Though XML offers its users many advantages like simplicity, extensibility, interoperability; information retrieval from XML document is very difficult task. So database research field concentrates on XML as a database. User need to know structure of the document before querying the document to know the semantics which require forming query. XML documents are flexible and do not have fixed schema, so user may fail to retrieve information as answer to query. Frequent patterns of XML documents provide intentional knowledge of the document and they specify information of the document in terms of a set of properties instead of only set of data satisfying the query. Intentional answers are approximate and take less time. This knowledge is provided by XML mining tool which in terms of a set of tree based association rules. TAR provides rules in the form TB=>TH, where TB is body tree and TH is the head tree of the rule and TB is a subtree of TH. These rules are helpful for the users to get implicit information about the document and thus it will be more useful for the system in query formulation. The proposed XML query answering support framework is as shown in fig. 1. The purpose of this framework is to perform data mining on XML and obtain intentional knowledge. The intentional knowledge is also in the form of XML. This is nothing but rules with supports and confidence. In other words the result of data mining is TARs (Tree-based Association Rules).

A. Fundamental concepts

Given an XML document, we extract two types of TARs: • A TAR is a structure TAR (sTAR) iff, for each node N contained in SH , cH (n)= ⊥ , that is, no data value is present in sTARs, i.e. they provide information only on the structure of the document.• A TAR,SB ⇒SH ,is an instance TAR (iTAR) iff SH contains at least one node n such that cH (n) = ⊥ , that is, iTARs provide information both on the structure and on the data values contained in a document. Since TARs provide an approximate view of both the content and the structure of an XML document, (1) sTARs can be used as an approximate Data Guide of the original document, to help users formulate queries; (2)iTARscan be used to provide intentional, approximate answers to user queries. By observing s TARs users can guess the structure of an XML document, and thus use this approximate schema to formulate a query when no DTD or schema is available: as Data Guides, sTARs represent a concise structural summary of XML documents. Differently from Data Guides, sTARs do not show all possible paths in the XML document but only the frequent paths. In particular, for each fragment, its support determines how frequent the substructure is. This means that sTARs provide a simple path index which supports path matching and can be used for the optimization of the query process. An index for an XML dataset is a pre-defined structure whose performances maximized when the query matches exactly the designed structure. Therefore, the goal, when designing an index, is to make it as similar as possible to the most frequent queries. By contrast, iTARs give an idea about the type of content of the different nodes.

II. Association Rule Mining

Association rule mining, one of the most important and well researched techniques of data mining. It aims to extract interesting correlations, frequent patterns, associations or casual structures among sets of items in the transaction databases or other data repositories. Mining Association rules are widely used in various areas such as telecommunication networks, market and risk management, inventory control etc. Discovering recurrent patterns inside XML documents provides high quality knowledge about the document content: frequent patterns are in fact intensional information about the data contained in the document itself, that is, they specify the document in terms of a set of properties rather than by means of data. In particular, the idea of mining association rules to provide summarized representations of XML documents has been investigated in many proposals either by using languages (e.g. XQuery) and techniques developed in the XML context, or by implementing graphor tree-based algorithms. A proposal is proposed for mining and storing TARs (Tree-based Association Rules) as a means to represent intensional knowledge in native XML. Intuitively, a TAR represents intensional knowledge in the form SB → SH, where SB is the body tree and SH the head tree of the rule and SB is a sub tree of SH. The rule SB → SH states that, if the tree SB appears in an XML document D, it is likely that the wider, tree SH also appears in D. fig 1 shows that the sample xml document and its induced sub trees Figure 1: a) an example of XML document, b) its tree-based representation, and c) three induced subtrees The increasing amount of very large XML datasets available to casual users is a most challenging problem and calls for an appropriate support to anciently gather knowledge from these data. Data mining, already widely applied to extract frequent correlations of values from both structured and semi structured datasets, is the appropriate tool for knowledge elicitation.

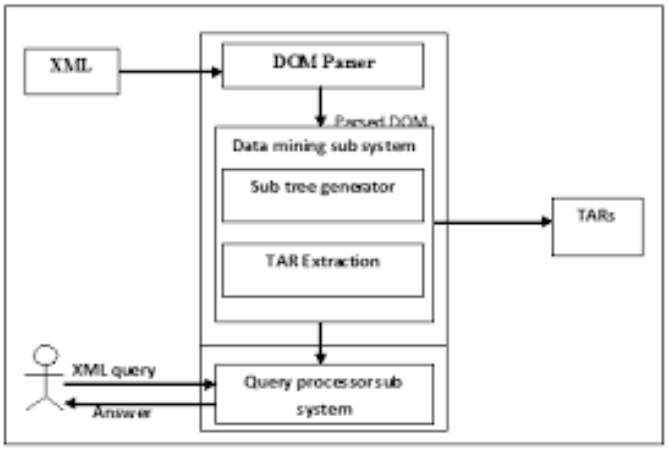

Fig 1: XML query answering support framework

A. Goals and Contribution:

The main goal is to provide a method for mining intensional knowledge from XML datasets by using tree based association rules. The mined TARs are basically used to get a concise idea of structure and the content of XML document. Also mined TARs are used for intensional query answering. The major advantage of our mining procedure is to directly work on the XML document without translating it into any intermediate format. And also the query is translated from original data set to TARs set. The paper contribution are : i) The use of CMTreeMiner for mining frequent subtrees from XML document. 2) Also translating user query into mined intensional knowledge. The aim of our proposed work is to use mined intensional knowledge instead of original document as well as to improve execution time of the queries over the mined intensional knowledge.

III. Extracting TAR's

Tree Based association rules are obtained by considering an items which having its support and confidence value above its user defined support and confidence from this a sub-trees are generated which having an aim to extract the collected data into the tree format so that data should be easily understood. This sub-trees having support and confidence from base tree this TAR's of two types, this can be seen in Fig 2 as below.

1) Content based (called Instance TAR's): This type of TAR's shows value or text in xml documents.
2) Structured TAR's: This type of TAR's shows structure of mined knowledge from xml documents.

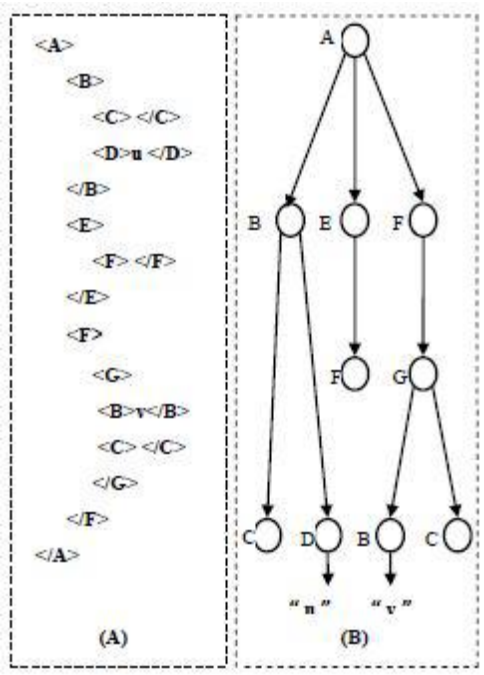

Fig 2. shows conversion of xml documents into tree based Association rules

Each and every rule is saved inside" <Rule>" element, which having three attributes ID, Support, Confidence

Algorithm 1. Get –Intersting-Rules (D,minisupp miniconf)
1: // To Search For frequent subtrees
2: FS=FindFrequent Subtrees (D, minisupp)
3: ruleSet = Ø
4: for all s FS do
5: // rules computed from s
6: tempSet = Compute-Rules (s, miniconf)
7: //For all rules
8: ruleset = ruleSet U tempSet
9: end for
10: return ruleSet

Function 1 Compute-Rules(s, miniconf)
1: ruleSet = Ø; blacklist = Ø
2: for all CS, subtrees of s do
3: if CSis not a subtree of any element in blacklist then
4: conf = sup(s) / sup(CS)
5: if conf >= minconf then
6: newRule = (CS,s,conf,supp(s))
7: ruleSet = ruleSet U {newRule}
8: else
9: blackList = blackListU CS
10: end if
11: end if
12: end for
13: return ruleSet

The following algorithms are obtained from[1][6][7][8] also algorithm 1 finds frequent sub trees and then hands each of them over to a function that computes all the possible rules. Depending on the number of frequent sub trees and their cardinality, the amount of rules generated by a naïve Compute Rules function may be very high.

A. Assigning Index to TAR's

TAR's provides intentional answer to the query which is more concise. Instead of describing data in terms of properties it gives the properties which data frequently satisfies. Index is assign to each path present in at least one rule. Index file is an XML document containing set of references to the each node in the rules.

IV. Query Answering

Once rule files and index files are saved as an dataset to be queried, user enter the query which is on original document. This query is transformed into intentional query and then fired on extracted datasets. Due to this user get intentional answer to the query. This answer is precise and gives property which is frequently satisfied. Condition in the query is matched with the nodes in index file and references of rules are obtained. Rule IDs return from index file are accepted and only those rules are searched for answer. Thus answer is return from mined knowledge not from original document. It is also useful in absence of original document. Queries are of different types such as σ and Π, count queries, Top k queries which are frequently asked.

V. Rules Updating

XML documents on web go on changing. In case of documents which are frequently undergoes changes instead of creating new rules previously obtained rule and index files are updated. This is done by creating dummy node while generating frequent subtree. Dummy nodes are the nodes whose confidence is less than minconf but it is near to that. When document changes and confidence value becomes greater than or equal to minconf only that respective dummy node gets activated and rule as well as indexes are updated.

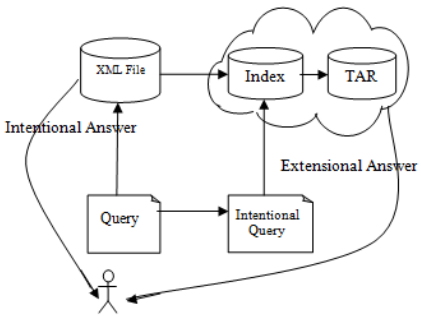

Fig.3 . System Architecture

VI. Conclusion and Future scope

With the growing importance of XML in document representation, new processing and integration technologies are being devised. The focus of this chapter, however, has been to describe, in general, the capability and benefits of data mining techniques in the XML domain and to conceptualize the XML mining process. This chapter attempts to show the improved knowledge discovery of both structure and content of XML documents with utilisation of data mining techniques. This explicitly expresses the representation of XML data and the broad categories of XML mining: XML structure mining and XML content mining. These categories are presented according to data mining tasks such as classification, clustering and association. Then presents the process of knowledge discovery from XML documents summarising the three tasks of clustering, association mining and classification on structure or/and content of XML documents. Further discussed the evolution of knowledge discovery where the current application of XML enables a simplified data mining process and makes the discovered patterns interchangeable among conforming data mining tools and other analytical applications. Then the protocols that support XML and data mining, making data mining possible across the web using XML. XML data mining is a challenging and exciting field with further possibilities. Following are some of the areas identified for future development**:**

Integration of XML Mining The integration of XML, the database languages, such as SQL, and data mining techniques will increase the functionality of relational database products and XML products. It will provide more user friendly mining. The larger RDBMS and data warehouse companies have already expressed an interest in integrating data mining and XML data models into their database products.

Graphical user interface Full integration of data mining products with other application tools and the use of GUIs will enhance usability. To satisfy the range of data mining users (from naive to expert users), future work should include mining user graphs that is structural information of web usages**,** as well as visualization of mined data using systems such as WWWPal system (*WWWPAL*).

Multimedia XML data To perform web content mining, keyword information and content for each of the nodes is required. This information will allow the automatic development of a set of keywords to distinguish text document, multimedia document or other kinds of document based on the contained characteristics such as color, brightness and texture. Data mining is able to intelligently prepare data and allow types of information to be distinguished

Security and Privacy As data mining is applied to large semantic documents or XML documents, extraction of information should consider privacy and rights management of shared data. XML mining should have the authorization level to empower security to restrict only to appropriate users to discover classified information.

References

[1] C R. Agrawal and R. Srikant. Fast algorithms for mining association rules in large databases. In Proc. of the 20th Int.Conf. on

[2] J.W.W. Wan and G.Dobbie, “Extraction of Association rules from XML Documents Using XQuery parser,” Proc. Fifth ACM Int Workshop Web Information and Data Management, pp.95-97, 2003.

[3] M. Mazuran, E. Quintarelli, and L. Tanca. Miningtree-based association rules from xml documents. Technical report, Politecnico

[4] Mirjana Mazuran, Elisa Quintarelli, and Letizia tanca. Optimized Data Mining for XML query-answering support.IEEE Transactions on Knowledge Data Engineering, Volume:PP Issue:99, 2011.

[5] G. Marchionini, Exploratory Search: From Finding to Understanding, Comm. ACM, vol. 49, no. 4, pp. 41-46, 2006.

[6] T. Asai, H. Arimura, T. Uno, and S. Nakano. Discovering of frequent substructures in large disordered trees. In TechnicalReportDOITR216,DepartmentofInformatics,Kyushuuniversity.http://www.i.kyushuu.ac.jp/doitr/trcs216.pdf,2 003

[7] A. Termier, M. Rousset, and M. Sebag. Dryade: “A new optimized approach for discovering closed frequent trees in heterogeneous tree databases”. In Proc. of the 4th IEEE Int. Conference. On knowledge and Data Mining, pages 544–548.

[8] K. Wang and H. Liu. Discovering typical structures of documents: a road map approaches for XML query answering support. In Proc. of the 21st Int. Conf. on Research and Development in Intensional Information Retrieval, pages 145–154, 1998

[9] S. Gasparini and E. Quintarelli, ―Intensional Query Answering to XQuery Expressions,‖ Proceeding. 16th International Conference on Database and Expert Systems Applications, Pp. 544-553, 2005.

[10] D. Braga, A. Campi, S. Ceri, M. Klemettinen, and P. Lanzi. Discovering interesting information in xml data with association rules. In *Proc. Of the ACM Symposium on Applied Computing*, pages 450–454, 2003.

[11] K. Wong, J. X. Yu, and N. Tang. Answering xml queries using path based indexes: A survey. *World Wide Web*, 9(3):277–299, 2006.
[12] Y. Xiao, J. F. Yao, Z. Li, and M. H. Dunham. Efficient data mining for maximal frequent subtrees. In *Proc. of the 3rd IEEE Int. Conf. on Data Mining*, page 379. IEEE Computer Society, 2003.
[13] L. Feng, T. S. Dillon, H. Weigand, and E. Chang. An xml-enabled association rule framework. In *Proc. of the 14th Int.Conf. on Database and Expert Systems Applications*, pages 88–97, 2003.
[14] L. Feng, T.S. Dillon, H. Weigand, and E. Chang, "An XML-Enabled Association Rule Framework," Proc. 14th Int'l Conf.Database and Expert Systems Applications, pp. 88-97, 2003.
[15] C. Combi, B. Oliboni, and R. Rossato, Querying XML Documents by Using Association Rules

