Data Mining For Intensional Query Answering Using Tree Based Association Rules

[1]Neha H. Nasre, [2]Kapil Hande
[1]Student of MTech second year, CSE Dept, Priyadarshini Bhagwati College of Engineering, Nagpur, India
[2]Assistant Professor, Dept. of CSE, Priyadarshini Bhagwati College of Engineering, Nagpur, India

***Abstract*- The database research field has concentrated on the Extensible Markup Language (XML) due to its flexible hierarchical nature which can use to represent huge amounts of data, also it doesn't have absolute and fixed schema, but having possibly irregular and incomplete structure. Extracting information from semi-structured documents is a very hard task, and is going to become more and more critical as the amount of digital information available on the internet grows. Indeed, documents are often so large that the dataset returned as answer to a query may be too big to convey interpretable knowledge. Since query languages for semi structured data rely the one document structure to convey its semantics, in order for query formulation to be effective users need to know this structure in advance, which is often not the case. When users specify queries without knowing the document structure, they may fail to retrieve information which was there, but under a different structure. An approach based on Tree- Based Association Rules (TARs), which provide approximate, intentional information about the structure and the contents of XML documents both, as well as it can be stored in XML format. This mined knowledge is used to provide, a concise idea of both the structure and the content of the XML document and quick, approximate answers to queries whenever required.**

***Keywords*- Extensible markup Language (XML), approximate query answering, data mining, intentional Information, Tree-Based Association Rules (TARs).**

1. INTRODUCTION

The data over the internet is not structured. It is thus not very easy to store and parse the stored data using databases. The database research field has concentrated on the Extensible Markup Language (XML) as a flexible hierarchical model suitable to represent huge amounts of data with no absolute and fixed schema, and a possibly irregular and incomplete structure.XML is used to represent huge amount of data without any absolute schema and structure. To retrieve information from XML document two techniques are used keyword search and query retrieval. Keyword search is used when we have to match exact desired word. Query retrieval is used whenever document is following certain schema but its availability of documents with schema is partially fulfilled. So when we search the desired query over document when we are unknown of the schema it fails.

Unstructured document causes excess of information to be included in answer which is not essential and formulation of query becomes difficult. If at all your formulation of query goes wrong the resultant system will thus fail to give exact expected answer. Mining of XML documents is quite different from structured data mining and text mining. The structure of an XML document is indicated by element tags and their nesting. It allows the representation of semi-structured and hierarchal data containing the values of individual items and the relationships between data items. Mining of contents along with structure provides new means into the process of knowledge discovery.

The idea of mining association rules to provide summarized representations of XML documents has been focused in many proposals either by using languages (e.g., XQuery) and techniques developed in the XML context, or by implementing graph- or tree-based algorithms. The technique of mining and storing Tree-Based Association Rules (TARs) as a means to represent intentional knowledge in original XML is used here. TARs are extracted for two main purposes to get a little idea of both the structure and content of an XML document and to use them for intentional query-answering, which allows the user to query the extracted TARs rather than the original document.

2. REVIEW OF RELATED WORKS

There are number of researches available in the literature for XML document mining to efficiently store, index and search on XML data. In the recent years, several researchers are focused on mining based encoding the XML nodes. The following explanation gives some of the recent researches to mine information from the XML documents. Apriori [10] algorithm and AprioriTid [10] algorithm uses multiple passes over data present in the databases for finding the frequent pattern from the data. The support values of each data were counted and the items were determined which have minimum support count given by the user. In succeeding passes, it proceeds with items passed in the previous passes. The items were stored into new large itemsets are called candidate itemsets. This process continues until no large itemsets were retrieved.

XMINE RULE [8] operator is used for mining association rules with the relational data. This intends that, after dropping of unneeded data, XML document is converted into relational form. XQuery along with Apriori [1], [7] extracts association rules

and Query Answering system made by XQuery used only in simple XML documents. XML document can be mined using XQuery language without pre-processing or post-processing for association rules.

XPATH [5], the most important component of XQuery was used in the field of mining association rules. It allows the user to specify the rules. It retrieves the answers for the queries made. Frequent subtree mining [5] technique was used in the discovery process which does not ignore the structure of the tree in the rules. The frequent subtree is splitted into several subtrees based on the value of support provided by the user.

DRYADEPARENT [11], tree mining algorithm that illustrates the depth of the frequent patterns to identify and branching factor were the key factors. It retrieves the embedded subtrees that sustain the relationships between the node pairs. But it does not distinguish the pairs. DataGuides [12] provides very short and correct formulation of queries in semi structured databases. This gives the usefulness of schemas in semi structured databases. Where there is no specific or fixed schema. By the schemas this method contributes the query formulation to the user and query optimization to the query processor.

3. EXISTING APPROACH

There is no existing approach has yet studied the problem of relevance oriented result ranking in depth. The search intention for a keyword based query is not easy to determine and can be equivocal, because the search through condition is not unique; hence, to measure the confidence of each search intention candidate, and to rank the individual matches of all these candidates is a challenging task. Subsisting methods cannot resolve this ranking strategy to rank the individual matches challenge, thus it return low quality result in term of query relevance. Disadvantages of Existing System: Search intention for a keyword query is not easy to determine. It returns low result quality in term of query relevance. Rank the individual matches of all these queries are challenging.

4. PROPOSED APPROACH

The proposed XML query answering support framework is to perform data mining on XML and obtain intentional knowledge. The intentional knowledge mined is also in the form of XML. This is nothing but rules with support and confidence. In other words, the result of data mined is TARs(Tree-based Association Rules).

As can be seen in fig. 4.1, the framework is to have data mining for XML query answering support. When XML file is given as input to the DOM parser, it will parse until it is well formed and validness. If the XML document is valid, it is parsed and loaded into a DOM object 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 sub system. This module takes XML query from the end user and makes use of mined knowledge to answer the query quickly.

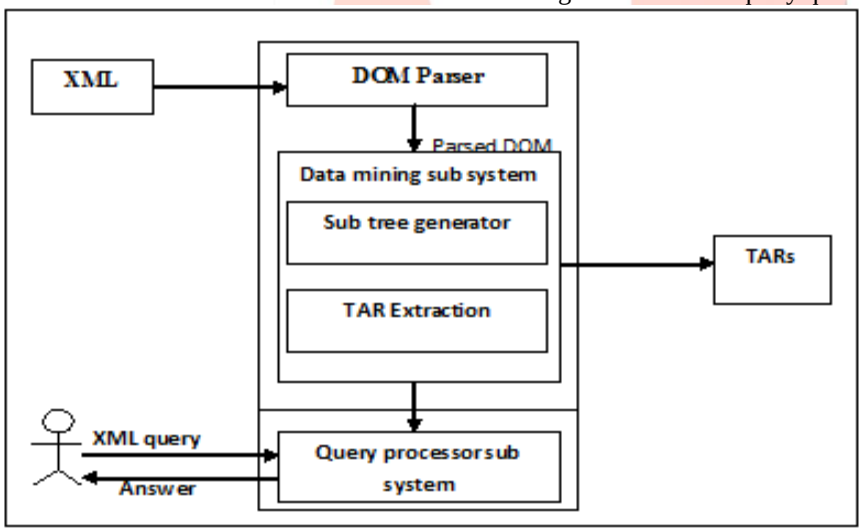

Fig 4.1: XML query answering support framework

5. TAR EXTRACTION

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 5.1

5.1) Content based (called Instance TAR's): This type of TAR's shows value or text in xml documents.

5.2) Structured TAR's: This type of TAR's shows structure of mined knowledge from xml documents.

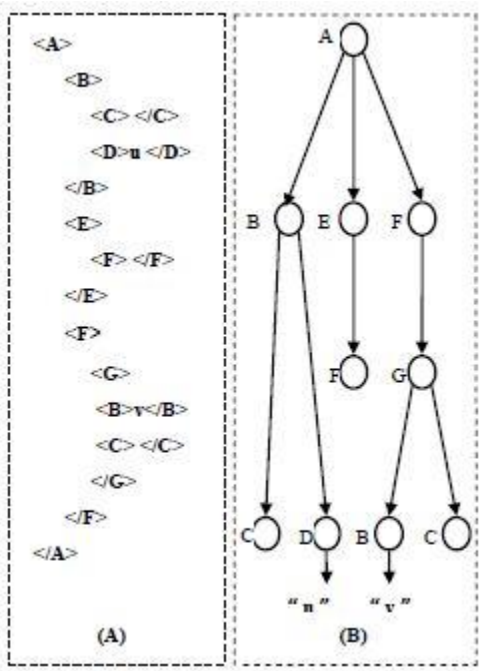

Fig 5.1 shows conversion of xml documents into tree based Association rule

6. MODIFICATIONS OF XML DOCUMENT

The answers were retrieved for the given XQuery over the entire XML document which gets transformed to TAR files. When the original XML document gets modified, the TAR file has been updated by using XPATH. This is used to update the XML files dynamically instead of creating the new TAR file for the modified XML document. This also updates the index files already created. This simplifies the work when the input XML document changes at any time. Also by using with XML DOM parser the TAR files and its index can be updated.

7. TREE RULER PROTOTYPE

TreeRuler [1] is a tool used in our approach. When the XML document is given, it makes users to retrieve the intensional information for the queries. Users formulate XQueries over the original data, and these queries are automatically translated and executed on the intensional knowledge. Get the Gist allows intensional information extraction from an XML document, when given the supports, confidence and the files where the extracted TARs and their index are to be stored.

The tree ruler interface offers three tabs : Get the Idea: this allows showing the intensional information as well as the original document, to give users the possibilities to compare the two kinds of information. Get the Answers: it allows querying the intensional knowledge and the original XML document.

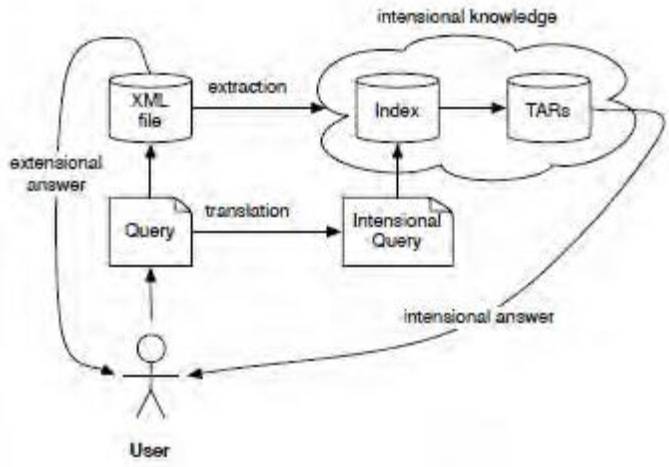

Fig 7.1 Tree ruler architecture

Users have to write an extensional query. When the query belongs to the classes have analyzed, then it is translated and applied to the intensional knowledge. Finally, once the query is executed, the TARs that reflect the search criteria are shown in Fig.7.1

8. CONCLUSION

Conventional database answers to queries, usually given as lists of objects, are not always the best means of efficient and effective communications between users and the database systems, especially when the answers include a large number of objects. Our method is for deriving intentional knowledge from XML documents in the form of TARs, and then storing these TARs as an alternative, synthetic data set to be queried for providing quick and summarized answers. This procedure has characterized by some key aspects that, it works directly on the XML documents, without transforming the data into any intermediate format. It looks for general association rules, without the need to impose what should be contained in the antecedent and consequent of the rule. It stores association rules in XML format and it translates the queries on the original data set into queries on the TARs set. The aim of this project is to provide a way to use intentional knowledge as a Substitute of the original document during querying and not to improve the execution time of the queries over the original XML data set.

9. REFERENCES

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

[2] Alfiya Iqbal Ahmed Shaikh, Sanchika Bajpai, "Frequent Pattern Mining for XML Query-Answering Support", International Journal of Innovative Technology and Exploring Engineering, ISSN: 2278-3075, Volume-4 Issue-2, July 2014.

[3] Swati N. Patil and R. V. Mane, "Tree Based Association Rules Mined From Xml Document For Xml Query Answering", International Journal of Advanced Computational Engineering and Networking, ISSN: 2320-2106, Volume-2, Issue-5, May-2014.

[4] V. Kasthuri Muthu, A. Sameera T. "Mining Algorithm for XML Query Answering Support", ijcset, vol4, 2013.

[5] J. Paik, H.Y. Youn, and U.M. Kim, "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, 2005

[6] Y. Chi, Y. Yang, Y. Xia, and R. R. Muntz, "Cmtreeminer: Mining both closed and maximal frequent subtrees", In Proc. of the 8th Pacific-Asia Conf. on Knowledge Discovery and Data Mining, pages 63–73, 2004

[7] J.W.W. Wan and G. Dobbie, "Extracting Association Rules from XML Documents Using XQuery," Proc. Fifth ACM Int'l Workshop Web Information and Data Management, pp. 94-97, 2003.

[8] Braga, A. Campi, S. Ceri, M. Klemettinen, and P. Lanzi, "Discovering Interesting Information in XML Data with Association Rules," Proc. ACM Symp. Applied Computing, pp. 450-454, 2003.

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

[10] R. Agrawal and R. Srikant, "Answering XML Queries by Means of Data Summaries", Proc. 20th international Conf. Very Large Data Bases, pp. 478- 499, 1994.

[11] A. Termier, M.Rousset, M.Sebag, K.Ohara, T.Washio, and H.Motoda, "DryadeParent: An Effective, efficient and Robust Closed Attribute algorithm for tree Mining," IEEE Transaction on Data Mining., vol. 20, Pg: 301-321, Mar. 2008.

[12] R. Goldman and J. Widom, "DataGuides: Enabling Query Formulations and Optimization techniques in Semi structured Databases," Proc. 23rd Int Conf. on Very Large Data Bases.