ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Website: www.ijircce.com

Vol. 5, Issue 5, May 2017

Data Mining Study and Analysis for Semi Structured Data Extraction Using XML

Versha[1], Deepika Garg[2]

M. Tech, Dept. of CSE, Advanced Institute of Technology and Management, Palwal, India[1]

Assistant Professor, Dept. of CSE, Advanced Institute of Technology and Management, Palwal, India[2]

ABSTRACT: It becomes very difficult to extract intentional knowledge from semi structured data.As per the study of different researches going to extract the useful knowledge from unconstructed environment is very useful to meet the desired output. For those requirements we need to use TAR called Tree based association rule.Every research has its own outcome but we are giving the best suitable mixture of the study about the semi structured data extraction using xml. In this paper representing the combination of various studies about the data mining approach using xml query i.e very efficient and quickly responsive approach. We are providing the various consolidated result for data extraction using xml data mining technique.

KEYWORDS: TAR; xml query; data mining ; data extraction

I. INTRODUCTION

XML has become a popular format for storing and sharing data across heterogeneous platforms. It is widely used in applications as it can allow applications to have communication though they are built in different platforms. The XML documents are plenty in enterprises and the data retrieval can be done in two ways. The first approach is that user gives keywords and the program searches for relevant documents. The second approach is give XML queries that are answered. In this paper we use query-answering system to access XML documents. The users are supposed to post queries and this system is much efficient to make questions and retrieve answers respectively. Discovering recurrent patterns inside XML documents provide high-quality knowledge about the document content: frequent patterns are in fact intentional information about the data, that they specify the document in terms of a set of properties rather than by means of data.

This paper addresses the need of getting the gist of the document before querying it, both in terms of content and structure. Discovering recurrent patterns inside XML documents provides high-quality knowledge about the document content: frequent patterns are in fact intentional 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. As opposed to the detailed and precise information conveyed by the data, this information is partial and often approximate, but synthetic, and concerns both the document structure and its content.When users specify queries without knowing the document structure, they may fail to retrieve information which was there, but under a different structure. This limitation is a crucial problem which did not emerge in the context of relational database management systems. Frequent, dramatic outcomes of this situation are either the information overload problem, where too much data are included in the answer because the set of keywords specified for the search captures too many meanings, or the information deprivation problem, where either the use of inappropriate keywords, or the wrong formulation of the query, prevent the user from receiving the correct answer.

II. RELATED WORK

In [1] authors provides a method for deriving intentional knowledge from XML documents in the form of TARs, and then storing these TARs as an alternative, synthetic dataset to be queried for providing quick and summarized answers.

ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Website: www.ijircce.com

Vol. 5, Issue 5, May 2017

Our procedure is characterized by the following key aspects: a) it works directly on the XML documents, without transforming the data into any intermediate format, b) it looks for general association rules, without the need to impose what should be contained in the antecedent and consequent of the rule, c) it stores association rules in XML format, and d) it translates the queries on the original dataset into queries on the TARs set..In [2] authors states that XML has become a defacto standard for storing, sharing and exchanging information across heterogeneous platforms. The XML content is growing day by day in rapid pace. Enterprises need to make queries on XML databases frequently. As huge XML data is available, it is challenging task to extract required data from XML database. It is computationally expensive to answer queries without any support. Towards this, in this paper we present a technique known as Tree-based Association Rules (TARs) mined rules that provide required information on structure and content of XML file and the TARs are also stored in XML format. The mined knowledge (TARs) used later for XML query answering support. This enables quick and accurate answering. We also developed a prototype application to demonstrate the efficiency of the proposed system. In [3]XMINE RULE [3] 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.

III. PROPOSED WORK AND FRAMEWORK

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). TAR mining is a the process composed of two steps: 1) mining frequent sub trees[1], which means sub trees with a support above a user-defined threshold, from XML document; 2)computing interesting rules, that is, rules with a confidence above a user-defined threshold , from the frequent sub trees. When the mining process has been finished and frequent TARs have been extracted, and are kept in XML format. This decision has been taken to allow the use of the same language (XQuery)[12] for analyzing both the original dataset and the finded rules. Association rules describe the frequent occurrence of data items in a large amount of data collected. A and B are the two data items. They are represented in the form of A∩B. Association rule is measured by means of Support and Confidence. Support represents the frequency of the set (A and B) found in the data set. Confidence represents the conditional probability of finding B, having got A. The interesting patterns among the subtrees of the given XML document can be identified. TAR mining is a process composed of two steps: 1) Mining frequent sub trees, that is, sub trees with a support above a user defined threshold value, from the XML document; 2) Computing interesting rules, defines finding the interesting rules that are with user-defined confidence value. The frequent pattern of subtrees had been extracted into TAR files. The TAR files are stored in XML document. These TAR files contain the rules which are computed over confidence values. Each rule is saved inside the element. These files represent the intentional knowledge about the XML document. This process of mining TAR eases the exploitation of the query-answering system. TARs are mined by generating the rules with the more number of nodes in the body tree.

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).

ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Website: www.ijircce.com

Vol. 5, Issue 5, May 2017

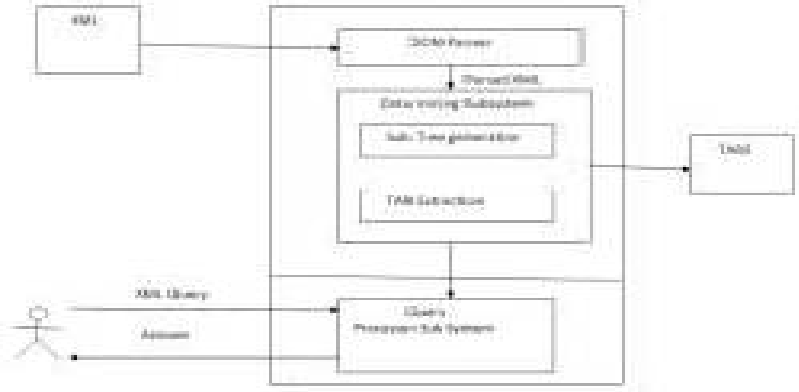

Fig. 1 – Proposed XML query answering support framework

As can be seen in fig. 1, the framework is to have data mining for XML query answering support. When XML file is given as input, DOM parser will parse it for well formedness and validness. If the given 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.

IV. PARAMETERIZED RESULTS

Extraction Time:

Extraction time depends on the number of nodes in xml document. Extraction time growth is almost linear with the respect to cardinality of the XML tree. Time required for the extraction of the intentional knowledge from in XML database. As no of nodes increases extraction time increase initially, it remains stable for sometimes and as no of nodes becomes too high again it increases very fast.

Answering Time:

Answer Time of getting intentional answer is comparatively less than that of extensional answers, as instead of accessing original document mined rule file is used to answer the query. Comparison with Support and Confidence Extraction time of generating rules from XML documents changes according to support and confidence . This can be show in graph by keeping first confidence constant and vary support and then keeping support constant.

Accuracy:

Accuracy of intentional answer is measured in terms of precision and recall. Query answering depends on support threshold. When support is high then chance of correct answering is high as less no of rules are to be access

V. CONCLUSION AND FUTURE WORK

TAR (Tree-based Association Rule) files from the original XML document which was given as input. This TAR files were formed by the frequent pattern of XML document. This TAR files were used by the query-answering system for retrieving the answers from the XML document. The query-answering system is used in the fields of information representation and reasoning, multimedia, sentimental analysis, information retrieval and natural language processing. The main goals we have achieved in this work are: 1) mine all frequent association rules without imposing any a-priori restriction on the structure and the content of the rules; 2) store mined information in XML format; 3) use extracted knowledge to gain information about the original datasets.

ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Website: www.ijircce.com

Vol. 5, Issue 5, May 2017

REFERENCES

1. R. Agrawal and R. Srikant. Fast algorithms for mining association rulesin large databases. In Proc. of the 20th Int. Conf. on Very Large DataBases, pages 487–499. Morgan Kaufmann Publishers Inc., 1994.
2. T. Asai, H. Arimura, T. Uno, and S. Nakano. Discovering frequentsubstructures in large unordered trees. In Technical Report DOITR216, Department of Informatics, Kyushu University. http://www.i.kyushuu.ac.jp/doitr/trcs216.pdf, 2003.
3. D. Barbosa, L. Mignet, and P. Veltri. Studying the xml web: Gatheringstatistics from an xml sample. World Wide Web, 8(4):413– 438, 2005.
4. Gary Marchionini. Exploratory search: from finding to understanding.Communications of the ACM, 49(4):41–46, 2006.
5. 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 [2] World Wide Web Consortium, Extensible Markup Language(XML) 1.0, http://www.w3C.org/TR/REC-xml/, 1998
6. D. Braga, A. Campi, S. Ceri, M.Klemettinen, and P. Lanzi, "Discovering of Interesting Information in XML Data with Association Rules," Proc. [4] R. Agrawal and R. Srikant. Fast algorithms for mining association rules in large databases. In Proc. of the 20th Int. Conf. on Very Large Data Bases. Morgan Kaufmann Publishers Inc., 1994.
7. 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.
8. J. Paik, H.Y. Youn, and U.M. Kim, "New Method for Mining Association Rules from a Collective XML Documents,"
9. 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.
10. 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.