INTERNATIONAL JOURNAL OF RESEARCH
In Advanced Engineering Technologies

Volume I, Issue 2 FEB 2014

DATA MINING FOR XML QUERY-ANSWERING SUPPORT

A.Ranjith Kumar1, M.venkata Krishna Reddy2

1M.Tech Student, Dept of CSE Jaya Prakash Narayan College of Engineering and Technology, Mahabubnagar, A.P, India

2Associate Professor, Dept of CSE, Jaya Prakash Narayan College of Engineering and Technology, Mahabubnagar, A.P, India

ABSTRACT-Extracting information from web 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. As the maintenance of several databases is a difficult task .In this paper we are storing our web documents in an XML format and is easy to search for a query .If we search for a query it starts from lowest common ancestor to its root node. We can call it as bottom to top approach. In this work we describe an approach based on Tree-based Association Rules (TARs) mined rules, which provide approximate, intentional information on both the structure and the contents of XML documents, and can be stored in XML format as well.

Keywords: Extensible Markup Language (XML), Treebased Association Rules (TARs)

1. INTRODUCTION

In recent years the database research field has concentrated on XML (extensible Markup Language) 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. There are two main approaches to XML document access: keywordbased search and query-answering. The first one comes from the tradition of information retrieval, where most searches are performed on the textual content of the document; this means that no advantage is derived from the semantics conveyed by the document structure As forquery-answering, since query languages for semi structured data rely the 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. In fact, it is not mandatory for an XML document to have a defined schema: 50% of the documents on the web do not possess one. 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.

As a consequence, 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 document 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.

BLOCK DIAGRAM

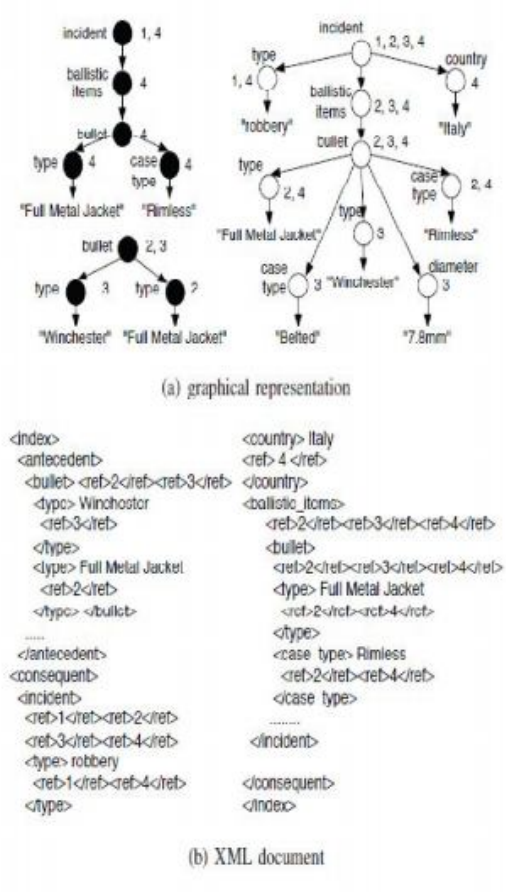

Figure 1. Shows the block diagram of TARS

II. METHODOLOGY

Admin: Admin maintains the total information about the entire application. Admin maintain the data in XML format only.User: user search queries and he got the reply in xml format. Xml Query Answering: In this project user search the information in semi structure document. He got reply in xml format only.

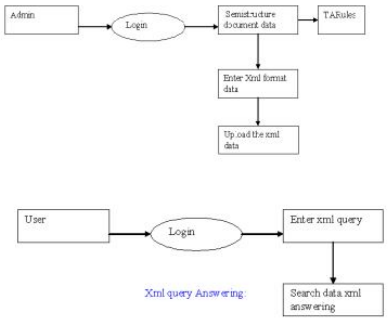

Fig 2. Query answering support.

Fig 2. Query answering support.Here the classification is implemented where the complete dividing of the entire data takes place. After the involvement of the division process depending on the set properties the effective classification of the data takes place. Where at the time of the classification the pertinent features are set which is similar to that of the array which is defined as the collection of the similar elements [10].

Therefore after the classification followed by the grouping takes place, where the elements of the single group are similar to one another that is considered as the one single cluster. So here the properties are set based on the method oriented with ensemble approach takes place. Therefore the present system is designed in an efficient fashion where the accurate data retrieval takes place respectively. Therefore the present designed method is displayed in the above figure which explains in an elaborative fashion oriented phenomena.

III. EXPECTED RESULTS

A lot of analysis has been made on the present method it completely overcome the drawback of the several existing techniques respectively. Where it isapplied on the large number of the data sets and it is effective in terms of the classification based phenomena respectively.

A comparative analysis is made between the present method to that of the several previous existing method and it is also displayed on the below figure in the form of the graphical representation.

Therefore the present system is effective and efficient in terms of the performance based strategy and its accurate classification based phenomena respectively.

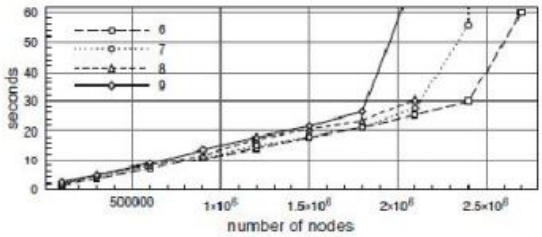

Fig 3. Extraction time with respect to Nodes

II. CONCLUSION

The main goals we have achieved in this work are:
1) Store mined information in XML format;
2) Use extracted knowledge to gain
3) Advantages are: Burden on database servers is overcome in this paper as well as an xml file format provides language interoperability i.e., reducing the burden on the programmer. Information about the original datasets .We have developed a C# prototype that has been used to test the effectiveness of our proposal. We have notdiscussed the updatability of both the document storing TARs and their index.

III. REFERENCES

[1] R. Agrawal and R. Srikant. Fast algorithms for mining association rules in large databases. In Proc. of the 20th Int. Conf. on Very Large DataBases, pages 487–499. Morgan Kaufmann Publishers Inc., 1994

[2] T. Asai, K. Abe, S. Kawasoe, H. Arimura, H. Sakamoto, and S. Arikawa. Efficient substructure discovery from large semi-structured data. In Proc.of the SIAM Int. Conf. on Data Mining, 2002.

[3] T. Asai, H. Arimura, T. Uno, and S. Nakano. Discovering frequent substructures in large unordered trees. In Technical Report DOI-TR 216, Department of Informatics, Kyushu University.

[4] E. Baralis, P. Garza, E. Quintarelli, and L. Tanca. Answering xml queriesby means of data summaries. ACM Transactions on Information Systems, 25(3):10, 2007.

[5] D. Barbosa, L. Mignet, and P. Veltri. Studying the xml web: Gathering statistics from an xml sample. World Wide Web,8(4):413–438, 2005.

[6] 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.

[7] 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.

8] C. Combi, B. Oliboni, and R. Rossato. Querying xml documents by using association rules. In Proc. of the 16th Int. Conf. on Database and Expert Systems Applications, pages 1020–1024, 2005.

[9] A. Evfimievski, R. Srikant, R. Agrawal, and J. Gehrke. Privacy preserving mining of association rules. In Proc. of the 8th ACM Int Conf. on Knowledge Discovery and DatMining, pages 217–228, 2002.

[10] 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.

Authors

Mr. A. RANJITH KUMAR doing his M. Tech in CSE from JAYA PRAKASH NARAYAN COLLEGE OF ENGINEERING, MAHABUBNAGAR. His Interested areas of Research include data mining.

Mr. M.VENKATA KRISHNA REDDY is working as associate professor (CSE) in JAYA PRAKASH NARAYAN COLLEGE OF ENGINEERING, MAHABUBNAGAR. His Interested areas of Research include data mining.