International Journal of Contemporary Research in Computer Science and Technology (IJCRCST) e-ISSN: 2395-5325
Volume2, Issue 6 (June '2016)

ADVANCED DATA MINING FOR XML QUERY ANSWERING SUPPORT

N.Saritha,
Department of Computer Science and Engineering,
Sasurie Academy of Engineering,
Coimbatore-641653, India.

M.Aruna,
Assistant Professor,
Department of Computer Science and Engineering,
Sasurie Academy of Engineering,
Coimbatore-641653, India.

Abstract: 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. 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. 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, (ii) quick, approximate answers to queries and (iii) output without redundancy, null tags and empty tag. In this work we focus on the second and third feature. A prototype system and experimental results demonstrate the effectiveness of the approach.

Keywords: XML, approximate query-answering, data mining, intentional information.

I.INTRODUCTION

Data mining is used for mining data from databases and finding out meaningful patterns from the database. Many organizations are now using these data mining techniques. This review of literature focuses on how data mining techniques are used for different application areas for finding out meaningful pattern from the database.

For better decision making, the large repositories data collected from different resources require proper mechanism of extracting knowledge from the databases. Knowledge discovery in databases (KDD), often called data mining, extracting information and patterns from data in large data base. The core functionalities of data mining are applying various techniques to identify nuggets of information of decision making knowledge in bodies of data. From the last decades, data mining and knowledge discovery applications have important significance in decision making and it has become an essential component in various organizations and fields. The field of data mining has been increased day by day in the areas of human life with various integrations and advancements in the fields of Statistics, Databases, Machine Learning, Pattern Reorganization, Artificial Intelligence and Computation capabilities etc.

XML has become a popular format for storing and sharing data across heterogeneous platforms. The XML format is neutral, flexible and interoperable. 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. The first approach is done using conventional information retrieval technique that works on the search process based on the given search word. With respect to query answering, it is not easy to process such request. To make this searching easy this paper presents data mining for XML query answering support. XML documents are validated by either DTD or schema. However, schema presence is not mandatory to process XML file. This paper presents data mining framework for XML query answering support. The XML documents essence is extracted and kept in another XML file in the form of TARs. With the help of this XML query answering becomes easy.

a) XQuery:

XQuery was designed to query XML data. XQuery is designed to query XML data - not just XML files, but anything that can appear as XML, including databases. XQuery[2] is built on XPath expressions. XQuery is a language for finding and extracting elements and attributes from XML documents. XQuery can be used to (a) Extract information to use in a Web Service. (b) Generate summary reports. (c) Transform XML data to XHTML. (d) Search Web documents for relevant information.

b) XQuery EXAMPLE

for $x inoc("student.xml")/college/stud where $x/rollno>30 order by $x/rollno return $x/name .

XQuery is compatible with several W3C standards, such as XML, Namespaces, XSLT, XPath, and XML Schema.

c) Tree-Based Association Rules from Xml Document

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)[4] 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.

d) 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.

II.EXISTING SYSTEM

There is no existing approach has studied the problem of relevance oriented result ranking in depth yet. The search intention for a keyword query is not easy to determine and can be ambiguous, because the search via condition is not unique; so, how to measure the confidence of each search intention candidate, and rank the individual matches of all these candidates are challenging. Existing methods cannot resolve this ranking strategy to rank the individual matches challenge, thus it return low result quality in term of query relevance.

Drawbacks

- 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

III. PROPOSED SCHEME

- Mine all frequent association rules without imposing any a-priori restriction on the structure and the content of the rules.
- Store mined information in XML format.
- Use extracted knowledge to gain information about the original datasets.

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 for query-answering, since query languages for semi structured data rely the on 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.

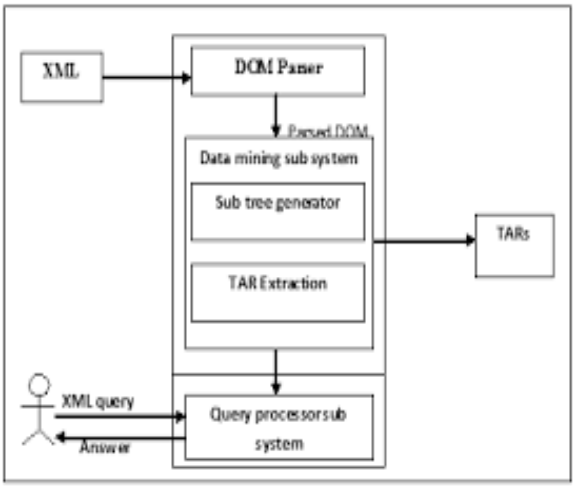


Fig 1: XML query answering support framework

a) The tree-ruler prototype:

TreeRuler is a tool that integrates the functionalities proposed in our approach. Given an XML document, it enables users to extract intensional knowledge and compose traditional queries as well as queries over the intensional knowledge, receiving both extensional and intensional answers. Users formulate XQueries over the original data, and queries are automatically translated and executed on the intensional knowledge. Fig 3 shows the tree ruler architecture.

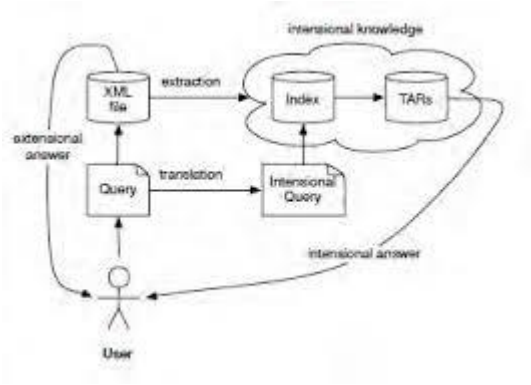


Figure 1: Tree Ruler Architecture

b) Advantages of proposed system:

- Resolve keyword ambiguity Problems.
- To effectively identify the type of target node, i.e. search for node.
- To effectively infer the types of condition nodes, i.e. search via node.
- Rank the individual matches of all possible search intentions.

IV. RESULTS AND DISCUSSION

Since data content in the internet is increasing day by day the retrieval of data is easy by using XML Query answering. The following screenshots shows a simple example of this.

- **Admin Login:**

Administrator can login with a predefined username and password. Authorized admin is the only person who can upload and create XML files.

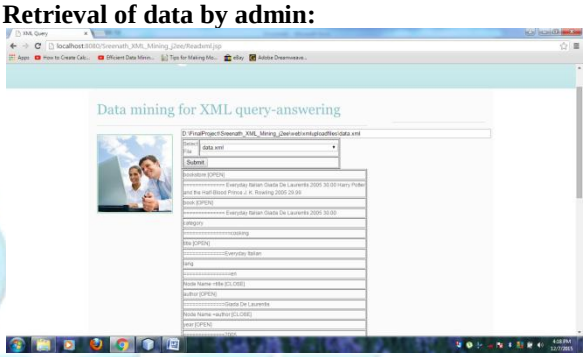
- **Admin Home:**

Admin is provided with the options to create, upload, delete xml files. The following shows the admin home page.

Uploading XML file:

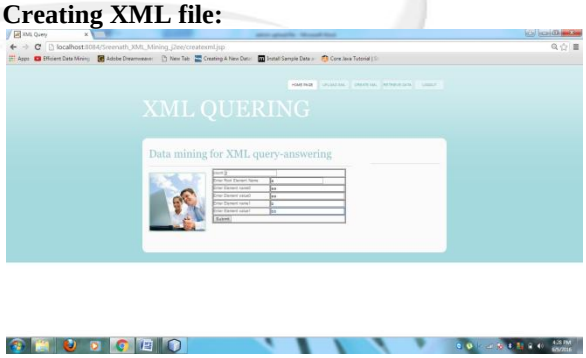
Admin can upload files from his computer.

Retrieval of data by admin:



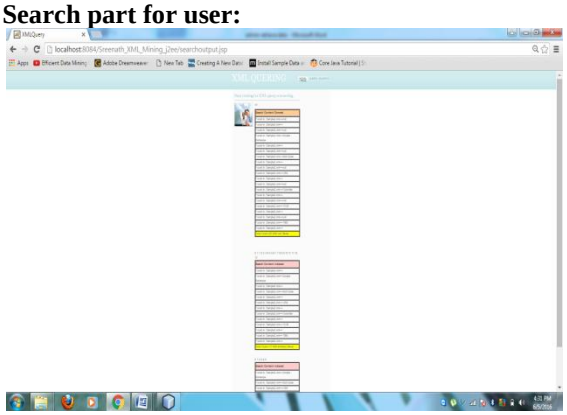
Admin can review the files that are uploaded.

Creating XML file:



Admin can create xml file as a stree structure by entering the number of levels of tree and the tree elements.

Search part for user:



User can search for the xml content for more appropriate data. This shows the refined levels of user output.

V.CONCLUSION

The main goals achieved in this work are:

- Mine all frequent association rules without imposing any a-priori restriction on the structure and the content of the rules;
- Store mined information in XML format;
- Use extracted knowledge to gain information about the original datasets.

We have not discussed the updatability of both the document storing TARs and their index .As an ongoing work, we are studying how to incrementally update mined TARs when the original XML datasets change and how to further optimize our mining algorithm; moreover, for the moment we deal with a (substantial) fragment of XQuery; we would like to find the exact fragment of XQuery which lends itself to translation into intensional queries.

VI.REFERENCES

[1] 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]http://ieeexplore.ieee.org/xpl/articleDetails.jsp?arnumber=6807687&sortType%3Dasc_p_Sequence%26filter%3DAND%28p_IS_Number%3A69459
35%29.

[3] G. Seshadri Sekhar1, Dr.S. Murali Krishna,\E_cient Data Mining for XML QASS/IOSR Journal of Computer Engineering (IOSRJCE)ISSN: 2278- 0661Volume 4, Issue 6 (Sep.-Oct. 2012), PP 13-22.

[4] KC. Ravi Kumar1, E. Krishnaveni Reddy2, Ramadevi.G3,\Data Mining for XML QASSt\,IOSR Journal of Computer Engineering (IOSR-JCE) ISSN:2278- 0661, ISBN: 2278-8727 Volume 5, Issue 6 (Sep-Oct. 2012), PP25-29.

[5] Univ. of Grenoble, St. Martin d'Heres.IEEE Transactions on Knowledge and Data Engineering (Impact Factor: 1.89). 09/2007; 20:300-320. DOI: 10.1109/TKDE.2007.190695

[6] World Wide Web Consortium. XML Schema, 2001. http://www.w3C.org/TR/xmlschema-1

[7]http://www.cs.washingtom.edu/research/xmldatasets/data/auctionsyahoo.xml

[8] World Wide Web Consortium. XQuery 1.0: An XML query language, 2007. http://www.w3C.org/TR/xquery.K. Elissa, "Title of paper if known," unpublished.