The International Journal Of Engineering And Science (IJES)
|| Volume || 3 || Issue || 7 || Pages || 24-30 || 2014 ||
ISSN (e): 2319 – 1813 ISSN (p): 2319 – 1805

Intentional Knowledge Mining Using Association Mining Technique in XML Query Answering Scheme

[1,]Ujwal Arjun Bodke , [2,]Santosh Kumar
[1,2,]*Department of Computer Science, SITRC*

----------ABSTRACT----------

Semistructured document's data retrieval has become a defacto standard for data storing, sharing and exchanging information over heterogeneous platforms. The XML content is growing speedily thus information withdral from semistructured documents has become very tedious task, companies or organizations need to make queries on XML databases frequently. As their XML data is huge, it is difficult task to mine intensional data from XML database. It is computationally costly to answer queries without any support. Thus in this paper we propose a technique known as Tree-based Association Rules (TARs) mined rules that extract required information on structure and content of XML file and the TARs are also stored in XML format provide a way to use intentional knowledge as a alternative of the original document during querying . This empower swift and precise respond for queries. We also built a framework to manifest the efficiency of the proposed system. The actual results are very constructive and query answering is expected to benificial in real time applications.

KEYWORDS : *XML dataset, Query Answering,TARS, Data Mining..*

Date of Submission: 22 June 2014 Date of Publication: 25 July 2014

I. INTRODUCTION

XML has become a favored format for data sharing and data storage across multiplex platforms. The XML format is neutral, workable and interoperable . It is universally used to communicate applications from heterogeneous platforms. The XML documents are sufficient in enterprises and the data retrieval can be done in two ways. The first approach is that user fire the XQuery on original XML document but this traditional approach is difficult and time consuming. So to overcome the problems of traditional approach we are introducing the new approach in which at thefirsttime retrieval of largedataset We will gain some commoninformationabout XML document(structural and semantic characteristics).This information helps analysis on more individual component.The necessity of getting the common meaning of the document before querying it, in terms of content and structure. XML finds recurrent arrangement inside XML documents which provide high-grade knowledge about the document content. Recurrent patterns are in fact intended information about the data contained in the document itself ,it means, they specify the document in terms of a set of belongings instead of by means of data. Instead of comprehensive and accurate information deliver by the data, this information is limited and often relative, but not genuine, and concerns both the document structure and its content.

II. PROPOSED FRAMEWORK

The framework shown in fig. 1 is proposed XML query answering support System . The purpose of this framework is to querying on intentional knowledge instead of original XML document. The calculated knowledge is also in the XML format. These are rules with supports and confidence. In other words the result of data mining is TARs (Tree-based Association Rules)

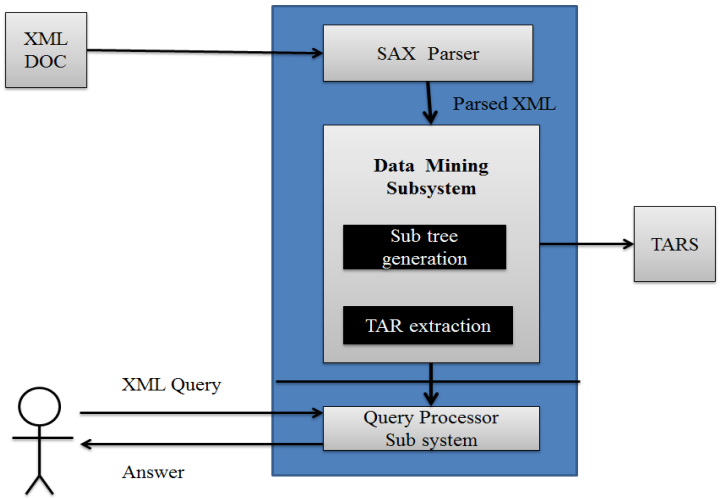

Fig : 1- Proposed XML query answering support framework

When XML file is given as input, SAX parser will resolve it for wellformedness and validness. If the given XML document is valid, it is parsed and loaded into a SAX object which can be operate conveniently. The parsed XML file is forwarded to data mining sub system in which two processes are done 1. sub tree generation 2. TAR extraction. The Query Processor Sub System uses generated TARs to answer the user's query. End user gives the query to this Query Processor Subsystem Module and this module gives the final accurate result in seconds, because this module uses Intentional data only(TAR extracted data) instead of original XML data.

Tar Extraction

It is basically two step process. In the first step recurrent patterns that satisfy given support are mined. In the second step interesting rules that have confidence above given threshold are determined from the frequent sub trees. Finding frequent sub trees is described in [1], [2], [6], [7], [8], [9]. Algorithm 1 finds frequent sub trees and calculates interesting rules.

Build Tree using SAX
Loading Xml file using SAX Builder

[1] Create Document object
[2] Create and Initialize SAXBuilder
[3] Build the xml file into SAXBuilder
[4] Assign the builder to Document object

Generating Tree Model
[1] Create DefaultMutableTreeNode object and assign the root element of Document object
[2] Iterate childrens of root element
[3] Process of every childrens of childrens and their attributes
[4] Add the childrens to DefaultMutableTreeNode object
[5] Iterate all child nodes until last node.

Build pathList using Pruning algo
[1] Create a object of ArrayList<PathListModel>
[2] Get the DefaultMutableTreeNode object of Tree Model
[3] Get the root node of Tree Model
[4] Recurse the Tree Model
[5] Get the current node
[6] Read the details of the node (name, depth, level, parent, etc...)
[7] Build path of parent node
[8] Create a PathListModel object and assign the values of current node

[9] Add PathListModel object to list of ArrayList<PathListModel> object
10. Iterate same for all child nodes

Depending on the number of recurrent sub trees and the cardinality, the amount of rules .The rules obtained are written to an XML file. Then indexing is made. Afterwards when XML queries are made, the proposed system uses index and TARs and quickly answers the query. Then indexing is made. Eventually when XML queries are made, the proposed system uses index and TARs and quickly answers the query.

III. XML QA SUPPORT SYSTEM

The exracted TAR files are used for querying rather the entire XML document. This reduce the extraction time as compare to querying the entire XML document. The rule item contains three aspect such as ID, Support and Confidence. An index [8] is created for the mined TAR file to make faster access of the document when queries are put and this index file is also created in XML format. This accommodate the set of trees and each node in the tree contains references to the generated rules. The query, which is made on the original XML document, automatically transfer on TAR files. By using this solution, the XML documents can be queried easily compared to the other operators [3]. This is because the XQuery, the XML query language which is specifically designed for XML documents. This consists of three class queries [8] to be transformed. They are as follows

Class 1:Node/child node queries:
This query is used to bring down simple and complex to count the number of elements having operators with restrictions on them.

Class 2: count-queries: This query is used specific data mentioned in the query.

Class 3: top-k queries: This query is used to select the top k queries which satisfy a grouping condition.

Class 4: Aggregate Queries: This query is used to apply aggregate functions(sum, min, max, avg).

Class 5: DML Queries: These queries used to manipulate the XML document by user without having any knowledge of XML .By using these class queries, the users can pose a query over the XML document.

IV. EXPERIMENTS AND RESULTS

Environment

The environment used to develop the template includes JSE (Java Standard Edition) 7.0,1 Net Beans IDE that run in Windows 7/8 OS. The Java API is used to build graphical user interface while IO and JAXP (Java API for XML Parsing) are used for implementing functionality. The main GUI has provision to choose XML file as input.It also allows choosing a file for storing extracted rules. Plain text view tab show the XML file content. Extended Graphical tree view tab shows the same XML file in tree format. XML node path list tab shows name of the node ,its path ,depth and level .

XML document Example

```
---------------------------------------------------------------------------------------------------------------------------
-----
 <articles>
<article>
<volume>30</volume>
<number>2</number>
<month>June</month>
<conference>ACM SIGMOD International Conference on Management of Data</conference>
<date>May 21 - 24, 2001</date>
<location>Santa Barbara, California, USA</location>
<title>Securing XML Documents ...</title>
<authors>
<author>E. Brown</author>
<author>L. Baines</author>
</authors>
```

<indexTerms>
<term>XML</term>
<term>Security</term>
<term>XQuery</term>
<term>Theory</term>
</indexTerms>
</article>
<article>
<volume>30</volume>
<number>2</number>
<month>June</month>
<conference>ACM SIGMOD International Conference on Management of Data</conference>
<date>May 21 - 24, 2001</date>
<location>Santa Barbara, California, USA</location>
<title>Securing XML Documents ...</title>
<authors>
<author>E. Brown</author>
<author>L. Baines</author>
</authors>
<indexTerms>
<term>XML</term>
<term>Security</term>
<term>XQuery</term>
<term>Theory</term>
</indexTerms>
</article>
<article>
<volume>30</volume>
<number>2</number>
<month>June</month>
<conference>ACM SIGMOD International Conference on Management of Data</conference>
<date>May 21 - 24, 2001</date>
<location>Santa Barbara, California, USA</location>
<title>Securing XML Documents ...</title>
<authors>
<author>E. Brownn</author>
<author>L. Baines</author>
</authors>
<indexTerms>
<term>XML</term>
<term>Security</term>
<term>XQuery</term>
<term>Theory</term>
</indexTerms>
</article>
<article>
<volume>30</volume>
<number>2</number>
<month>June</month>
<conference>ACM SIGMOD International Conference on Management of Data</conference>
<date>May 21 - 24, 2001</date>
<location>Santa Barbara, California, USA</location>
<title>Securing XML Documents ...</title>
<authors>
<author>E. Brown</author>
<author>L. Baines</author>
</authors>

<indexTerms>
<term>XML</term>
<term>Security</term>
<term>XQuery</term>
<term>Theory</term>
</indexTerms>
</article>
<article>
<volume>25</volume>
<number>6</number>
<month>June</month>
<conference>ACM SIGMOD International Conference on Management of Data</conference>
<date>June 4-6, 1996</date>
<location>Montreal, P. Q., Cananda</location>
<title>Query caching and optimization in ...</title>
<authors>
<author>B. Tannen</author>
<author>J. Parker</author>
<author>M. Strickland</author>
<author>B. Gale</author>
</authors>
<indexTerms>
<term>Measurement</term>
<term>Performance</term>
<term>Theory</term>
</indexTerms>
</article>
</articles>

Above is a sample XML document. We are doing experiments on this document. In this paper we are doing experiment on static data i.e. we are using ruleset file of [11] for association mining rules for the input XQuery on above sample XML document named as 'article'.

<ruleSet>
<AssociationRule support="0.3" confidence="0.8">
<RuleBody>
<item><ItemName>author</ItemName><ItemValue>E. Brown</ItemValue></item>
</RuleBody>
<RuleHead>
<item><ItemName>term</ItemName><ItemValue>XML</ItemValue></item>
</RuleHead>
</AssociationRule>
<AssociationRule support="0.3" confidence="0.5">
<RuleBody>
<item><ItemName>term</ItemName><ItemValue>Theory</ItemValue></item>
</RuleBody>
<RuleHead>
<item><ItemName>author</ItemName><ItemValue>B. Gale</ItemValue></item>
</RuleHead>
</AssociationRule>
<AssociationRule support="0.36" confidence="0.7">
<RuleBody>
<item><ItemName>author</ItemName><ItemValue>J. Parker</ItemValue></item>
</RuleBody>
<RuleHead>
<item><ItemName>term</ItemName><ItemValue>Measurement</ItemValue></item>

```
</RuleHead>
</AssociationRule>
<AssociationRule support="0.27" confidence="0.5">
<RuleBody>
<item><ItemName>conference</ItemName><ItemValue>ACM SIGMOD Internat...</ItemValue></item>
</RuleBody>
<RuleHead>
<item><ItemName>author</ItemName><ItemValue>E. Brown</ItemValue></item>
</RuleHead>
</AssociationRule>
<AssociationRule support="0.4" confidence="0.8">
<RuleBody>
<item><ItemName>year</ItemName><ItemValue>1996</ItemValue></item>
</RuleBody>
<RuleHead>
<item><ItemName>conference</ItemName><ItemValue>ACM SIGMOD Internat...</ItemValue></item>
</RuleHead>
</AssociationRule>
<AssociationRule support="0.4" confidence="0.74">
<RuleBody>
<item><ItemName>year</ItemName><ItemValue>2001</ItemValue></item>
</RuleBody>
<RuleHead>
<item><ItemName>conference</ItemName><ItemValue>ACM SIGMOD Internat...</ItemValue></item>
</RuleHead>
</AssociationRule>
<AssociationRule support="0.2" confidence="0.7">
<RuleBody>
<item><ItemName>term</ItemName><ItemValue>XML</ItemValue></item>
<item><ItemName>term</ItemName><ItemValue>XQuery</ItemValue></item>
</RuleBody>
<RuleHead>
<item><ItemName>author</ItemName><ItemValue>E. Brown</ItemValue></item>
</RuleHead>
</AssociationRule>
</ruleSet>
```

Above is a file[11] in which we have write extracted association rules. In which we got the Association rule in terms of support and confidence of our sample XML document .Also in this ruleset file we can see the RuleHead
(return value) of XQuery and RuleBody (condition)of the same XQuery.

V. RESULT

Querying over TAR files needs less time when compares to querying the original XML document. The extraction time was calculated for processing the intentional knowledge over the various numbers of nodes in the XML document. Precision and recall values were used to evaluate the accuracyof the results retrieved. In below result table we can see the performance difference of both the queries .Our proposed system is much faster than normal XQuery in execution.We have implement the queries mentioned in the [12]

Table 1.1 shows the output in milisecs.

Query	XQuery	Association rule based Query
Select query	40.4 msec	12.4msec
AND Query	18.5 msec	17.2 msec
OR Query	20 msec	15.2 msec
Count	16.6 msec	15.2 msec
Top-K	25 msec	15.5 msec

VI. CONCLUSION

In this paper we developed a java template for extracting TARs from given XML document so as to support XML queries. Thus the purpose of this paper is to extract recurrent pattern and store the extracted data in XML format; use the TARs to sustain query answering or to acquire information from XML databases. A framework is developed to test the efficiency of the proposed idea. The application introduce TAR's file by taking XML file as input and then finally index file that aid in query processing. The experimental results disclose that the proposed application is beneficial and can be used in real time applications.

VII. FUTURE SCOPE

We have performed these experiments of different queries on static data. But in the future we can generate RuleSet of input Xquery dynamically/automatically and also develop template for providing solution for most of the queries/operations on XML document.

REFERENCES

[1] 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.

[2] 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.

[3] Chandra Sekhar.K, 2Dhanasree,\Extracting TARs from XML for Efficient QA\International Journal of Computer Science and Network (IJCSN)Volume 1, Issue 6, December 2012.

[4] Mining Association Rules from XML Document using Modified Index Ta ble\,2013 International Conference on Computer Communication and Informatics(ICCCI -2013), Jan. 04 06, 2013.

[5] Anam V Bhaskara Reddy,\Answering Xml Query Using Tree Based Association Rules\,International Journal of Latest Trends in Engineering and Technology(IJLTET).

[6] D.Karthiga1, S.Gunasekaran,\Optimization of Query Processing in XML Document Using Association and Path Based Indexing\,International Journal of Innovative Research in Computer and communication Engineering Vol. 1, Issue 2,April 2013.

[7] Saranya T.J.,\MINING TREE-BASED ASSOCIATION RULES FROM XML DOCUMENT\,International Journal of Advanced Technology and Engineering Research (IJATER).

[8] Mrs.MopuriSujatha,\XML Query Answering using Tree based Association Rules\,International Journal of Advanced Research in Computer Engineering and Technology (IJARCET)Volume 1, Issue 8,October 2012.XML Query Answer System REFERANCES 21 [9] P. B. Vikhe1, B. L. Gunjal,\Extracting Tree Based Association Rules from XML Document\,International Journal,Volume 3, Issue 6, June 2013.

[10] Arundhati Birari1, Prof. RanjitGawande,\Mining Tree-Based Association Rules for XML Query Answering\,International Journal,Volume 2, Issue 3, May June 2013.

[11] Intensional Query Answering toXQuery expressionsSimone Gasparini and Elisa QuintarelliDipartimento di Elettronica e Informazione, Politecnico di Milano Piazza Leonardo da Vinci, 32 — 20133 Milano (Italy) fgasparini,quintarellig@elet.polimi.it

[12] XML Query-Answering Support System using Association Mining Technique I.Suganya1, N.Velmurugan2, Dr.P.Ganeshkumar31sugan.ilango@gmail.com, 2nvel_murugan@yahoo.com, 3drpganeshkumar@gmail.com

AUTHOR BIOGRAPHY

Miss.Ujwal Arjun Bodke She is post graduate student of computer engineering at SITRC Nashik under University of Pune. Her area of interest include Web mining.

Mr.Santosh Kumar is working as Assistant Professor in Computer department at SITRC,Nashik,Maharashtra,India.