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

International Journal of Innovative Research in Computer and Communication Engineering
Vol. 1, Issue 2, April 2013

Optimization of Query Processing in XML Document Using Association and Path Based Indexing

D.Karthiga[1], S.Gunasekaran[2]

Student,Dept. of CSE, V.S.B Engineering College, TamilNadu, India[1]

Assistant Professor,Dept. of CSE, V.S.B Engineering College, TamilNadu, India[2]

ABSTRACT: Data mining techniques extracts required information from the semi structured XML document. The association rule mining provides a tree representation for the XML document. In existing system the TAR (Tree Based Association Rule) would provide an approximate answering to the query, thus it decreases the query answering. Performance degradation is avoided by adding indexing technique to the existing technique, thus it increases the efficiency of the result. Classification introduces length analysis to provide accurate query answering to the users in less time.

Keywords:Approximate query-answering, Data Mining, Path Based Indexing,TAR,XML

I. INTRODUCTION

The XML (Extensible Markup Language) is now a days used in many aspects of web development, often to simplify data storage and sharing. It having various functionalities such as simplifies the data storage; platform changes and makes the data more available thus different applications can access our data. Because of these functionalities, the usage of XML document also grows higher in the organizations and companies. There are many techniques available to extract correct information from these documents. However, these techniques were not sufficient to retrieve an efficient answering from the semi-structured XML document. It leads to answering the query in a satisfactory manner as the semi-structured document may have irregular structure of document and redundant data's available. To recover that, an approach was introduced called the Tree Based on Association Rule [1], which would provide a Tree representation, which makes the users to get a less approximate detail to the query given. A fast retrieval to a query can be getting by the approach called Path Based Indexing mechanism. This indexing mechanism will help to visit the elements in a path manner thus; it increases the speed of the getting answer. The above mechanism would take time to get the answer for user query. Another technique called classification is applied in the TAR. This will help to reduce the complexity of calculating the weight and it give an exact answering.

II. XML

XML (Extensible Markup Language) is a flexible way to create common information formats and share both the format and the data on the World Wide Web, intranets, and elsewhere. XML can be used by any individual or group of individuals or companies that wants to share information in a consistent way. Thus there should be enhanced and new techniques to get exact information from the large amount of data residing in the internet or in a specified organization [1]. The XML format having a structured and a semi structured format in the usages. As the XML document growing in the internet may sometimes because the XML document to be a semi structured format. In the semi structured format the data retrieval will not be exact and efficient. Some of the data in the semi structured document will be null values and redundant values. This would create an unstructured format of XML document. For Example, the creation new Account in the Gmail consists of various mandatory and optional requirements. Consider, a new user fills all the mandatory details. The optional details are filled or leave empty according to the user's wish Thus, user left some of the optional details and these details are stored in an XML format. The above processes are happening in the front end. If the back end user wants to get the detail of any of optional requirements, there could be two possibilities: some of the user's data is absent and some of the data repeats. This kind of possibility leads to a semi-structured XML Document. It is a difficult process to formulate a query to get the answer from these semi structured document. This project provides a way to extract the answer from the query by using association rule mining and by indexing mechanism.

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

International Journal of Innovative Research in Computer and Communication Engineering
Vol. 1, Issue 2, April 2013

A. *Association Rule Mining*

Data Mining is a kind of efficient technique which would help to found information from the large amount of data in database. It allows users to analyse data from many different dimensions or angles, categorize it, and summarize the relationships identified. One of the mining techniques to found relationships between the data is Association rule mining. Association rules are created by analysing data for frequent patterns and using the criteria support and confidence to identify the most important relationships. Support is an indication of how frequently the items appear in the database. The rule $X \Rightarrow Y$ holds with support s if s% of transactions in D contains $X \cup Y$. Rules that have as greater than a user-specified support is said to have minimum support. For Example the support supp(X) of an itemset is defined as the proportion of transactions in the data set which contain the itemset. In the example database, the itemset {A,B,C} has a support of 1/5=0.2 since it occurs in 20% of all transactions (1 out of 5 transactions).Confidence indicates the number of times the if/then statements have been found to be true. The rule $X \Rightarrow Y$ holds with confidence c if c% of the transactions in D that contain X also contain Y. Rules that have a c greater than a user-specified confidence is said to have minimum confidence

B. Tree based on Association Rule

Association rules describe the co-occurrence of data items in a large amount of collected data and are represented as implications of the form X =>Y, where X and Y are two arbitrary sets of data items. The quality of an association rule is measured by means of support and confidenceEq (1) (2). In This paper the support and confidence is calculated by means of length analysis in XML document Fig (3). Thus it can be calculated as by following formula:

$$Support = \frac{\textbf{The searched Element length}}{\textbf{Total No of length in Document}} \quad (1)$$

$$Confidence = \frac{\textbf{The Searched Element}}{\textbf{Total No of length in Document}} \quad (2)$$

In this paper, we changed the formula of association rule introduced in the context of relational databases as per our project requirement to adapt it to the hierarchical nature of XML documents. The representation of an XML document as a tree (N,E,r,L,C) where N is the set of nodes, r is the root of the tree, E is the set of edges, L is the label function which returns the tag of nodes (with L the domain of all tags) and the content function which returns the content of nodes (with C the domain of all contents). In the Existing system, the TAR is used by element-only Infoset content model, where we added some functionality in the TAR of our proposed system and it retrieves the data by both element infoset and path based search.

C. Classification

Classification is a data mining function that assigns items in a collection to target categories or classes. The goal of classification is to accurately predict the target class for each case in the data. In XML the classification can be done by using the length analysis. The length of the document is classified using an execution file called Tree Ruler Classification. The analysed document will stored as XML format and applied in the further process to get the answering in a less time. A new data can be added in the classification, so that there is no problem for other processes. The algorithm for the length analysis is given below.

D. Experimentation

The relationship plays an important role in finding efficient way of answering query. We added some functionality in the TAR to get answer in a less time and less approximation manner. At First a pre-processing technique is introduced. In TAR a pre-processing technique Fig (1) is applied as Data Driven Approach. Thus a user can manually visit the elements in the XML document and can remove the unwanted elements. The pre-processed document is then applied to next process of weight calculation. The weight is calculated using the cosine range similarity value. The obtained values are clustered provide answer for the queries in the F.

$$Cosine\ Range\ Similarity = \frac{\mathbf{A.B}}{|\mathbf{A}|.|\mathbf{B}|} \quad (3)$$

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

International Journal of Innovative Research in Computer and Communication Engineering
Vol. 1, Issue 2, April 2013

The main process of our project is to make a tree representation to the given query of the user. The association rule makes a relationship between the elements to get the structure and content of XML document by length analysis. We calculated support and confidence value for each element [4]. From those values the tree can be constructed. The elements are placed in tree by these values.

E. Experimentation of Classification rule

There is an algorithm for the length analysis in the XML document which makes the query answering in an exact manner and in less time than the association rule mining.

1. Consider finding college details
2. Classify the data items from tree

   Ex: Class 1: Student

   Class 2: Staff

   Class 3: Department
3. User searching for details of a student

The length of the Tree will be predicted

Ex: Input: Student Name

The search will be: University->college name->department- >year->Student Name

4. User searching for details of a students in a department

The length of the Tree will be predicted

Ex: Input: Students in department

The search will be: University->college name->department->year->Total students

5. Time taken is reduced and Accuracy of query answering is obtained.

F. Query Answering

In this project we made a simple comparison between the normal query answering and the Tree representation answering. This answering can be done by XML Query (XQuery) [3] we classified the query answering as three kinds:

1) σ /n Queries

This is a kind of query where the AND and OR operators get used Fig (2). We combined AND and OR operators in this query answering method.

2) Count Queries

In this kind of queries, the total number of data present for that particular query will be the answer. If we want to calculate the number of names in the document means it can be obtained through this query.

3) Distinct-set:

The SQL DISTINCT clause allows you to remove duplicates from the result set. It can be used either in single field or in multiple fields. In single field, the simplest way to use the SQL DISTINCT clause would be to return a single field that removes the duplicates from the result set.

1II. INFORMATION RETRIEVAL

The information retrieval is the tracing and recovery of specific information from stored data. Information Retrieval and Data Mining are technologies for searching, analysing and organizing text documents like structured or semi structured data. In our project the retrieval can be done using the weight calculation with the help of cosine range similarity values. This similarity would analyse the distance between the two documents and would group the values. It will show how much percentage the values are in similar. Thus the similar values will be grouped. If the user given a query it would search on that group and retrieve the answer. The process of weight calculation would cause more time to get the

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

International Journal of Innovative Research in Computer and Communication Engineering
Vol. 1, Issue 2, April 2013

answer but the length analysis would retrieve the exact answering and in less time. The path based indexing is also the way of increasing the time of answering data.

IV. COMPARISON WITH TREE CLASSIFICATION

The above query answering methods were used in the normal way of answering methods after the clustering using weight calculation. The weight calculation provides a traditional way of query answering. This won't provide a full information for the semi structured XML document in accordance to the user's need. It may take time to search. For example if a user asking query to get the detail of a customer NAME and SALARY, the query would return the details accordingly. If the user further wants to get the detail such as NAME and DESIGNATION of the customer, the user again wants to give same query to get the detail. Thus to reduce the time constraints, a Tree Representation based on Association Rule would provide more information in a single tree Fig (5). The user doesn't need to repeat the query to get the various details about the same customer. Thus the availability of the query answering is increasing due to TAR representation. This process is based on the association rule mining. By using the classification rule's length analysis we can improve the accuracy and less time [5].

V. PATH BASED INDEXING

The application TAR is to view the each and every attribute in a tree representation manner. The indexing is a mechanism in which the value of index is stored and processed. The values are indexed by semi-TAR construction like personal, educational details and others. Thus after the tree construction the values of support and confidence can easily place in the corresponding places. So that, the searched item can searches through the indexed path values. In the existing work the indexing is the mechanism maintained by TAR itself which provide approximate answering to the query. But, because of using this Path Based Indexing [2] mechanism the approximation level is decreased and the exact answering to the query is increasing accordingly. In this project the path based indexing uses the two levels namely the content level and the ratio level Fig (4). The content level specifies the element present in the documents. In this project there are 100 documents present in the dataset. Each document is displayed in the content level with the elements. The ratio range for an element specifies the ratio of the corresponding element present in whole document. If the same element present in document is called redundancy. The redundancy can be obtained by using this ratio range.

VI.RESULTS

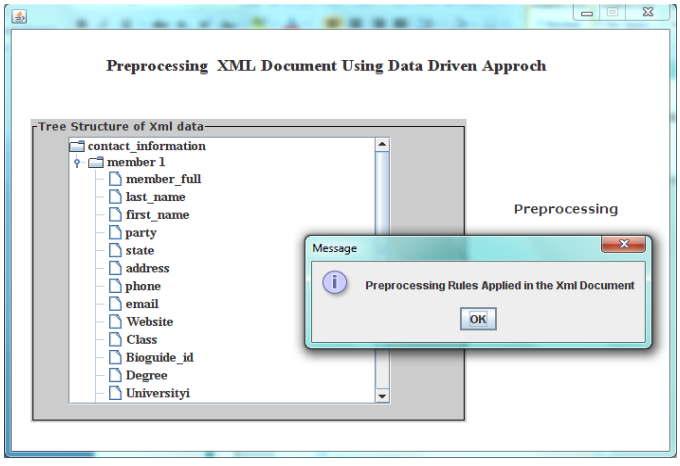

Fig 1.Preprocessing of an XML Document

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

International Journal of Innovative Research in Computer and Communication Engineering
Vol. 1, Issue 2, April 2013

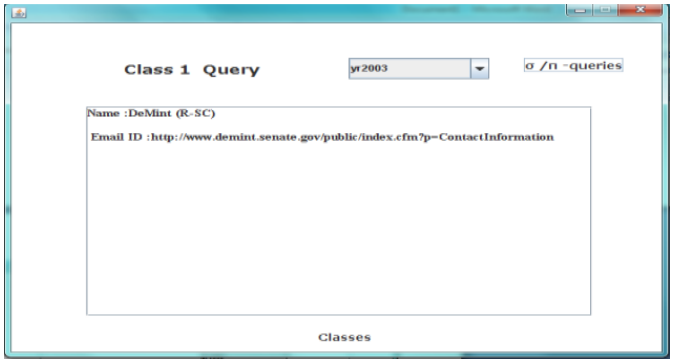

Fig 2.Traditional answering to the query after weight calculation

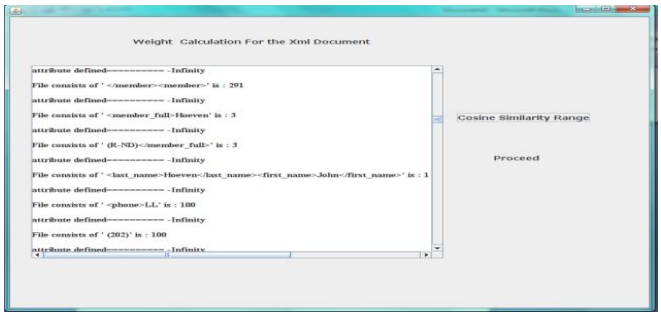

Fig 3.Calculation of support and confidence

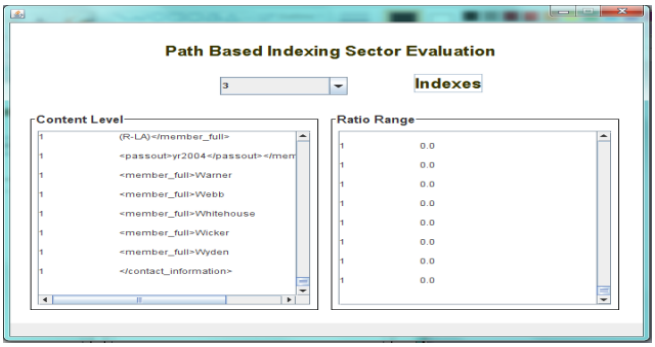

Fig 4.Path Based Indexing Mechanism

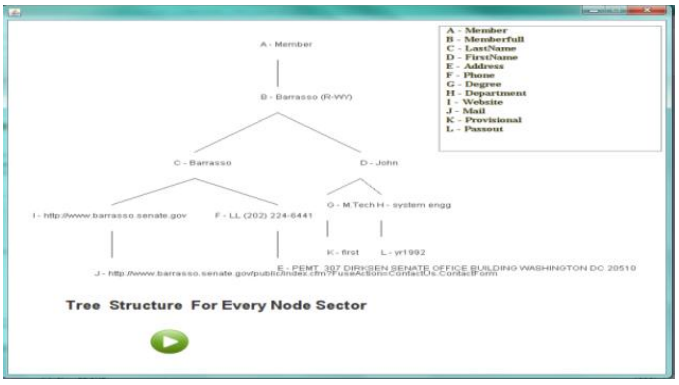

Fig 5.TAR View

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

International Journal of Innovative Research in Computer and Communication Engineering
Vol. 1, Issue 2, April 2013

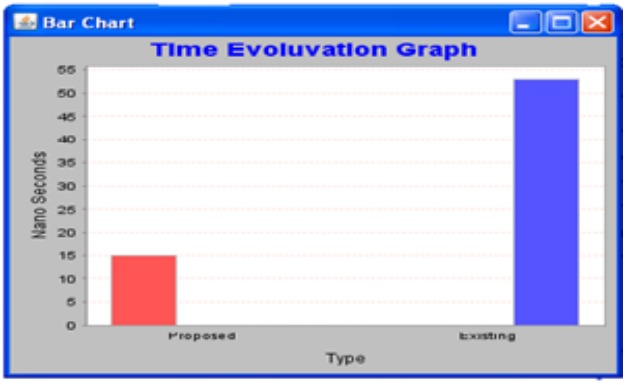

Fig 6.Comparision chart for Time Evaluation

VII. CONCLUSION

Mined all frequent association rules without imposing any restriction on the structure and the content of the rules. The proposed algorithm extends Path Based Indexing and allows users to extract efficient answering from XML documents. The main goals we have achieved are: 1) Mined frequent association rules gives the structure and the content of the XML file using tree representation [1]; 2) Stored mined information in XML format; as a consequence, 3) It can effectively use the extracted knowledge to gain information, by using query languages for XML, about the original datasets where the mining algorithm has been applied [4]. The exact information in TARs provides a valid support in several cases. It allows obtaining and storing implicit knowledge of the documents. When compared to the Association rule the classification would increases the efficiency of query answering and time reduction in searching a document. For any kind of XML document the user can easily get the accurate answering.

REFERENCES

[1]. MirjanaMazuran, ElisaQuintarelli, and LetiziaTanca"Data Mining for XML Query-Answering Support" IEEE TRANSACTIONS ON KNOWLEDGE AND DATA ENGINEERING L. 24, NO. 8, AUGUST 2013.

[2]. K. Wong, J.X. Yu, and N. Tang, "Answering XML Queries Using Indexes: A Survey," World Wide Web, vol. 9, no. 3, pp. 277-299, 2006.

[3]. 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.

[4]. C. Combi, B. Oliboni, and R. Rossato, "Querying XML Documents by Using Association Rules," Proc. 16th Int'l Conf. Database and Expert Systems Applications, pp. 1020-1024, 2005

[5]. Albert Bifet and RicardGavald "Adaptive XML Tree Classification on evolving data streams"UniversitatPolitecnica de Catalunya, Barcelona,Spain,2010