Exploiting Hierarchal Structure of XML Data Using Association Rule Analysis

Gurpreet Kaur and Naveen Aggarwal

***Abstract*—Data mining is the process of extracting useful information from the huge amount of data stored in the databases. Data mining tools and techniques help to predict business trends those can occur in near future. Association rule mining is an important technique to discover hidden relationships among items in the transaction. Association rules is a popular and well researched method for finding interesting relation between variables in large databases. For generating strong association rules, it depends on the association rule extraction by any algorithm for example Apriori algorithm or FP-growth etc and the evolution of the rules by different interestingness measure for example support/confidence, lift/interest, Correlation Coefficient, Statistical Correlation, Leverage, Conviction etc. The classical model of association rules mining is support-confidence. The goal is to experimentally evaluate association rule mining approaches in the context of XML databases. Algorithms are implemented using Java. For experimental evaluation different XML datasets are used. Apriori and FP Tree algorithm have been implemented and their performance is evaluated extensively.**

***Index terms*—Data mining, association rule analysis, XML.**

I. Introduction

Data mining or Knowledge discovery in databases (KDD) is the process of discovering previously unknown and “useful” patterns form the huge amount of data stored in flat files, databases, data warehouses or any other type of information repository. Database mining deals with the data stored in database management systems. Data mining tools can answer business questions that traditionally were too time consuming to resolve. There are basically two most important reasons that data mining has attracted a great deal of attention in the recent years [2]. First, our capability to collect and store the huge amount of data is rapidly increasing day by day. Due to the decrease in the cost of storage devices and increase in the processing power of computers, now a days it is possible to store huge amount of organizational data and process it. The second but the more important reason is the need to turn such data into useful information and knowledge.

If we are rich in data then we may or may not be rich in information, because the useful information is often hidden in the data. Data mining tools and techniques are used to generate information from the data that we have stored in our database repositories over the years. To take advantage in the market over the competitors, decision makers, administrators or managers need to mine the knowledge hidden in the data collected over the years and use that information in an effective and systematic way [2].

Data mining scans the databases for hidden patterns, finding predictive information that experts may miss because it lies outside their expectations. It consists of an iterative sequence of the following steps such as Data Cleaning, Data Integration, Data Selection, Data Transformation, Data Mining, Pattern Evaluation, Knowledge Extraction. The data mining step may interact with the user or a knowledge base. The interesting patterns are presented to the user and may be stored as new knowledge in the knowledge base.

A. Evolution of Data Mining

Evolution of data mining techniques began when business data was first stored on computers.

Improvements in data access techniques continued and technologies that allow users to navigate through their data in real time are also available. Data mining takes this process to a new dimension of data access and navigation to information delivery. Data mining is used as an application in the business community because it is supported by three technologies that are following [3]: Data collection, Multiprocessor computers, Data mining techniques and algorithms. Commercial databases are growing at unprecedented rates at different industries; market places etc. In the evolution from business data to information, each new step has built upon the previous one. Following are the four steps that allow business queries to be answered correctly [3]: Data Collection (1960s), Data Access (1980s), Data Warehousing & Decision Support (1990s) and Data Mining (Emerging Today).

B. Application Area of Data Mining

Some application areas are as follows:

- A pharmaceutical company [3] can analyze its recent sales and their results to improve targeting of high-value physicians and determine which marketing activities will have the greatest impact in the next few months.
- A credit card company [3] can access its large warehouse of customer transaction data to identify customers most likely to be interested in a new credit product etc

II. Association Rule Analysis: Apriori and FP-Tree Algorithm

Association rule mining is an interesting data mining technique that is used to find out interesting patterns or associations among the data items stored in the database. Support and confidence are two measures of the interestingness for the mined patterns. These are user supplied parameters and differ from user to user. Association rule mining is mainly used in market basket analysis, retail

Manuscript received October 5th, 2011, revised May 31, 2012.

Gurpreet Kaur is with the Chandigarh Group of Colleges, Gharuan, Mohali, India (email:gurpreetkaur7885@gmail.com).

Naveen Aggarwal is with the UIET Punjab University, Chandigarh, India (email: navagg@gmail.com).

data analysis, churn analysis and prevention, medical diagnosis and research or website navigation analysis. In market basket analysis we identify different buying habits of customers and analyze them to find associations among items those are purchased by customers. Items that are frequently purchased together by customers can be identified. Association analysis is used to help retailers to plan different types of marketing, item placement and inventory management strategies [1]. When we do association rule mining in relational database management systems we generally transform the database into (tid, item) format, where tid stands for transaction ID and item stands for different items purchased by the customers. There will be multiple entries for a given transaction ID, because one transaction ID indicates purchase of one particular customer and a customer can purchase as many items as he want. An association rule can look like this:

X (buys, Laptop) → X (buys, Windows 7 CD)
[support =1%, confidence=50%]
Where:

$$\text{Support} = \frac{\text{The number of transactions that contain Laptop and Windows 7 CD}}{\text{The total number of transactions}} \quad (1)$$

$$\text{Confidence} = \frac{\text{The number of transactions that contain Windows 7 CD}}{\text{The number of transactions that contain Laptop}} \quad (2)$$

The above rule will hold if its support and confidence are equal to or greater than the user specified minimum support and confidence. Because we are searching for frequent itemsets in the sample, it is possible that we may miss some global frequent itemsets. To lessen this we use lower support than minimum support for the sample. In this way we trade off some degree of accuracy against efficiency. There are various mechanisms so that we can find out all the missing frequent itemsets those are not find out in the sample.

A. Apriori Algorithm

The Apriori algorithm[1][5] is one of the most important algorithms for association rule mining because most of the other algorithms are based on it or extensions of it. It is a Main-memory based algorithm. Main memory imposes a limitation on the size of the dataset that can be mined. It states "Any subset of a frequent itemset must be frequent".

The algorithm executes in two steps as described above. i.e. frequent itemsets generation and association rule generation. The frequent itemsets generation is again a two step process [4]:

- Candidate itemsets (Ck) generation i.e. all possible combination of items those are potential candidates for frequent itemsets.
- Frequent itemsets (Fk) generation i.e. support for all candidate itemsets are generated and itemsets having support greater than the user-specified minimum support are qualified as the frequent itemsets.

This algorithm scans over the database a multiple number of times and it is not possible to find out number of scans earlier. Firstly Apriori algorithm generates frequent 1-itemsets F1 by directly reading the database D.Then it iterates through for loop and Fk-1 is used to generate candidate itemsets Ck. In the next pass Ck is then used in the generation of Fk. The ***generate*** procedure generates potential candidate itemsets and then eliminates itemsets from this set whose subset is not frequent. The algorithm builds a special hash tree data structure in the memory for support counting [1].

TABLE I: EXAMPLE OF DATABASE BROWSER

Browser Users	Browsers Used
A	IE, Mozilla, Netscape
B	IE, Mozilla
D	Netscape, Opera, Chrome
C	Mozilla, Safari

For rule A→ C
Support=Support ({AUB}) = 50%
Confidence= Support ({AUB})/Support ({A}) = 66.6%

TABLE II: EXAMPLE SHOWING SUPPORT FOR EACH BROWSER

Frequent Itemsets	Support
Mozilla	75%
Netscape	50%
IE	50%
{Mozilla, IE}	50%

B. FP-Tree Algorithm

FP-Growth is an algorithm for generating frequent item sets for association rules. This algorithm compresses a large database into a compact, frequent pattern– tree (FP tree) structure. FP-tree structure stores all necessary information about frequent itemsets in a database [2][6].

A frequent pattern tree (or FP-tree in short) is defined as:-

- The root labeled with "null" and set of items as the children of the root.
- Each node contains of three fields: item-name (holds the frequent item), count (number of transactions that share that node), and node- link (next node in the FP-tree).
- Frequent-item header table contains two fields, item-name and head of node link (points to the first node in the FP-tree holding the item).

III. IMPLEMENTATION DETAILS

XML has made it possible to improve its presentation and redefine the way in which documents and data were exchanged. As web is migrating from HTML to XML, large amount of data is accumulating day by day. This huge amount of data on the websites is needed to be managed. For the same purpose, many data mining techniques are available to manage the datasets. XML data has the levels maintained which are used to define a single attribute specifically i.e. it can provide the information about any single attribute at different levels. For example, if an address of any student is defined using an attribute address, then, at first level, country is defined, and then state, then city, then street number and at last house number. Therefore, on exploiting the hierarchy in any XML file about any attribute can further refine the results of any search query. The websites which have been built using XML can be used as the datasets by viewing their source. Moreover, there are many XML repositories available on internet which provides huge XML datasets.

These datasets can be used for the research work. After the selection of the XML datasets, Pre-processing and transformation of the dataset are done. During the transformation steps, Conversion of XML to CSV (Comma Separated Values) and Conversion of XML to ARFF (Attribute Relational File Format) is done. Therefore, the datasets selected are converted to CSV (Comma Separated Values) or ARFF (Attribute Relational File Format). Conversion to these file formats makes it easy to use XML datasets. One can exploit XML hierarchy levels using these file formats. An ARFF (Attribute-Relational File Format) file is an ASCII text file that describes a list of instances The CSV file is used and is stored in the database.

XML files can also be stored directly to the database at different levels. An XML document along with its associated schema is input into an XML parser. The parser checks that the document is well formed and, if the schema is also available, checks that the XML is valid according to what has been defined in the schema[9]. Because the schema is also an XML document, it is validated recursively against another schema, respectively. The parser then provides access methods for another application to access the data that was contained within the original XML document [5] sharing a set of attributes. ARFF files have two distinct sections. The first section is the Header information, which is followed the Data information. The Header of the ARFF file contains the name of the relation, a

list of the attributes (the columns in the data), and their types.

A. Organization of XML into Levels

The hierarchal University database [7] is shown below in Fig. 1.

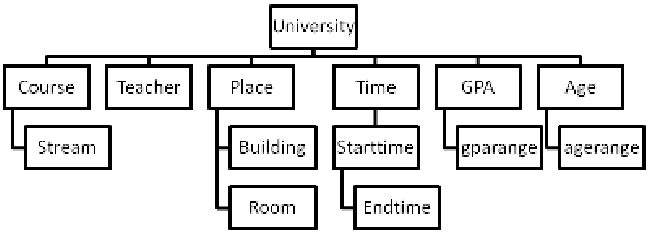

Fig. 1. Hierarchal structure of University Database.

Similarly hierarchy is maintained in other downloaded XML files and level wise data is stored in databases. In hierarchy we recursively apply association rule algorithm on each level in databases. After that steps followed during the implementation are listed in Fig. 2.

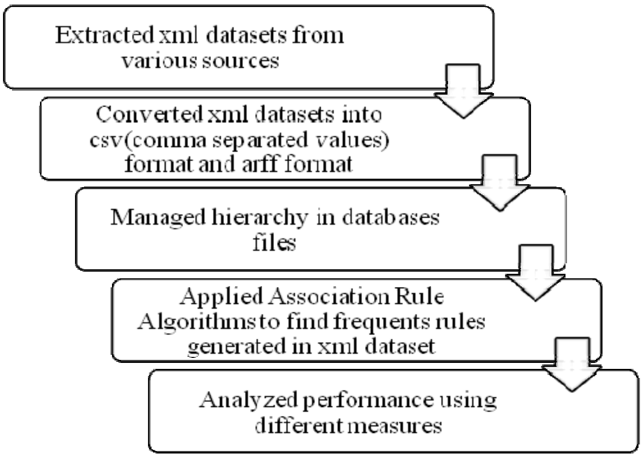
Extracted xml datasets from various sources
Converted xml datasets into csv(comma separated values) format and arff format
Managed hierarchy in databases files
Applied Association Rule Algorithms to find frequents rules generated in xml dataset
Analyzed performance using different measures

Fig. 2. Steps followed during implementation

Different algorithms (Apriori and FP Tree) are applied on the different XML datasets at different levels to analyze the performance of each algorithm at different levels. There performance can be analyzed on the basis of many parameters i.e. Support, Confidence, Execution time etc. Apriori and FP Tree algorithm techniques are used to exploit the hierarchical structure of XML data.

- When two different algorithms are applied on University [5] dataset, following graph has been obtained. It has been predicted by the graph that Apriori and FP-Tree generates same number of rules for each level in four different datasets. Therefore, as one moves down in the hierarchy and be more specific, results are better than the above levels.
  Minimum support=10%
  Minimum confidence=80%

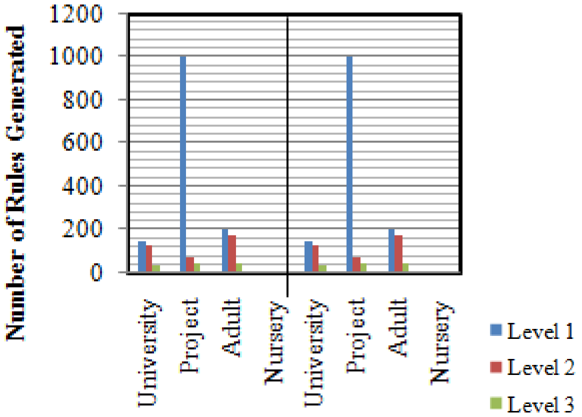

Fig. 3. Number of rules generated when Apriori and FP Tree applied on different datasets

- Performance Analysis of Execution time taken by Apriori and FP-Tree on each hierarchal level for selected four datasets downloaded from XML Repository[7][8].
  Minimum support=10%
  Minimum confidence=80%

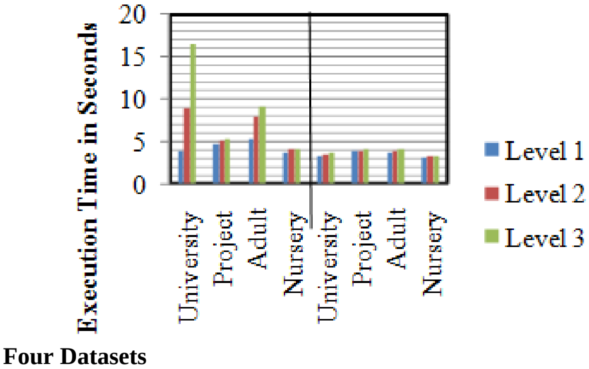

Fig. 4. Analysis of execution time when Apriori and FP Tree applied on different datasets.

It has been predicted that FP-Tree takes less time in generating rules as compared to Apriori at each level of four datasets.

When all association rule algorithms(Apriori, FP-Tree, Generalized Sequential Pattern[GSP], Predictive Apriori and Tertius) are applied on four datasets, Predictive Apriori generates maximum number of rules at minimum support and at maximum confidence as compared to all other algorithms, the sequence of generating max rules by algorithms are(Predictive Apriori > Tertius > Apriori and FP-Tree >

GSP).

Performance Analysis on Multilevel Association Analysis.

- At 1st Level Execution Time=3.753 secs, rules generated are 145.
- At 2nd Level Execution Time=8.804 sec, rules generated are 126.
- At 3rd Level Execution Time=16.499 sec, rules generated are 40.

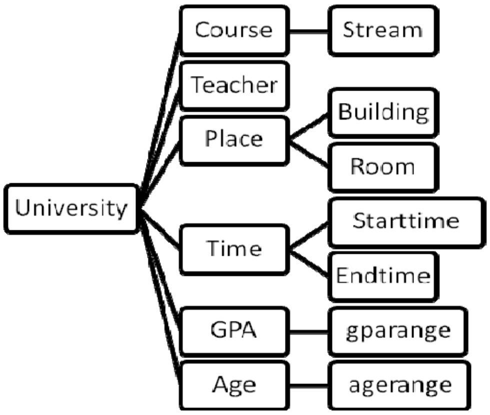

Fig. 5. Multilevel Association Rule Analyses.

Fig. 5. shows as we move down in a hierarchy the execution time increases, and the rules generated are more relevant, specific and decreases. In hierarchal XML databases, record details are more specific, relevant and understanding of databases increases. According to the client requirement one can decide whether more or less specific rules are required or not.

IV. Conclusion

XML has come a long way since it emerged, and it has been constantly improved and is still undergoing a lot of changes as the internet technology is growing[6]. But one factor which puts XML in an advantageous position is that it is through the changes that it has been flexible, and has been able to fit all the growing needs and functions as needed. XML is the foundation for many data formats, including HTML, WML, XHTML, and more. It has recently become popular because it can facilitate the transfer of data between widely disparate programs, operating systems, and companies. The key to XML's utility is that it enables any developer to design her own data format using her own terms and requirements. In fact, XML is so popular that Microsoft has built its entire suite of products, from operating systems to server components, around the concept of XML[8].

The real power of XML comes from the fact that with XML, not only can you define your own set of tags, but the rules specified by those tags need not be limited to formatting rules. XML allows you to define all sorts of tags with all sorts of rules, such as tags representing business rules or tags representing data description or data relationships.

In this we have discussed two algorithms (Apriori, FP-Tree) for association rule mining in the context of XML databases. We have discussed about the frequent itemsets, rule generation and time complexity on hierarchal XML databases. Rule generation is very simple as compared to frequent itemsets mining and it requires very less time. Extensive experiments have been performed to test the performance of these approaches over two real and one generated datasets

XML is fully compatible with applications like JAVA, and it can be combined with any application which is capable of processing XML irrespective of the platform it is being used on[8].

Frequent Pattern mining is used for finding frequent itemsets among items in a given data set. The results show that Apriori cannot be run very effective than FP -Tree. Apriori on the other hand runs too slow because each transaction contains density of each itemsets. By calculating the execution time we analyzed and concluded that both algorithm takes more time in mining frequent patterns from hierarchal datasets, as compared in mining flat datasets. When we move from top to bottom in a hierarchy, more detailed, simplified, specific, and less rules are generated as compared to non hierarchal dataset. When we analyse performance of all association rule algorithms on four selected datasets, Predictive Apriori generates maximum number of rules at minimum support and at maximum confidence as compared to other algorithms available for association analysis.

References

[1] M Kamber, J Han, "Data Mining: Concepts and Techniques", Second Edition, Morgan Kaufmann Publishers, 2000.

[2] Vipin Kumar, Tan Steinbach, "Data mining and techniques ", 2006.

[3] Alex Berson, J Smith Stephen, "Data Warehousing, Data Mining and OLAP".

[4] R. Agarwal, T.Imielinski and A. Swami, "Mining Association Rules between Sets of Items in Large Databases". In ACM SIGMOD International Conference in the Management of Data. Washington DC, 1993.

[5] J. Han, J. Pei and Y. Yin, "Mining Frequent Patterns without Candidate Generation". In ACM SIGMOD International Conference on Management of Data, Dallas, 2000.

[6] Qin Ding, Gnanasekaran Sundarraj, “Association Rule Mining from XML Data”, Computer Science Program 2006.

[7] Nigel Robinson, Mary Shapcott, ”Data Mining Information Visualization- Beyond Charts and Graphs”, 2002 IEEE.

[8] XMLDataRepository,http://www.cs.washington.edu/research/XMLdatasets/ [Online].

[9] McGovern, James, et al, "*Java 2 Enterprise Edition 1.4 Bible"*, Wiley, 2003, ISBN: 0764539663.

Gurpreet Kaur has received her B.Tech degree in Computer Science & Engineering from Chandigarh Engineering College, Landran, Mohali, India, in 2007 and M.E. Degree in Computer Science & Engineering from Panjab University, Chandigarh, India. Currently she is working as a Lecturer in Computer Science & Engineering at Chandigarh Group of Colleges, Gharuan, Mohali, India. She has more than 5 research publications in international and national conferences to her credit. Her research interests include Machine Learning, Association Rule Mining Techniques.

Naveen Aggarwal received his M.Tech degree in Computer Science & Engineering from IIT Kharagpur in 2002 and currently pursuing Ph. D at GGSIP University. He is working as Assisstant Professor at Panjab University, Chandigarh, INDIA. He has published more than 35 research papers in international journals and conferences. He is active member of Sun Developer’s Club and Computer Science Teachers Association. His main research interest includes Data Mining, Image Processing and Multimedia Computing.