Knowledge integration in a Parallel and distributed environment with association rule mining using XML data.

Mrs.Sujni Paul[1] [†] and Dr.V.Saravanan[††],
Ph.D Research Scholar Associate Professor & Head
Department of Computer Applications
Karunya University, Coimbatore, India.

Summary
In distributed data mining, the mining process is carried out in distributed locations parallel and generates frequent itemsets on the local areas. It is necessary to analyze these local patterns to gain global patterns when putting all the knowledge derived from local distributed location to a single one. Knowledge integration is the problem of combining the mined results obtained from the data residing at different sources, and providing the user with a unified view of these knowledge. Such a unified view is structured according to a so-called global schema, which represents the intentional level of the integrated data. Association rules are used for the mining process and hence local interestingness measure differs from the global interested patterns. Based on this criteria this paper focuses on the knowledge integration scheme from distributed workstations with XML data.

Key words:
Mining, Knowledge, parallel & distributed data mining, association rules, XML data.

1. Introduction

There exist many distributed locations today in a business or financial organizations. For example a large company may have subsidiary companies and each subsidiary company has its own data warehouse to make decisions for the development of the company. The decision taken will be in the distributed local workstations, which has to be met globally based on the interestingness measure. Data Mining and Knowledge Discovery in Databases (KDD) is an interdisciplinary field merging ideas from statistics, machine learning, databases, and parallel and distributed computing. It has been engendered by the phenomenal growth of data in all spheres of human endeavor, and the economic and scientific need to extract useful information from the collected data. The key challenge in data mining is the extraction of knowledge and insight from massive databases. Association rule mining, one of the most important and well researched techniques of data mining, was first introduced in [1]. It aims to extract interesting correlations, frequent patterns, associations or casual structures among sets of items in the transaction databases or other data repositories. Association rules are widely used in various areas such as telecommunication networks, market and risk management, inventory control etc. Association rule mining is to find out association rules that satisfy the predefined minimum support and confidence from a given data warehouse. The problem is usually decomposed into two subproblems. One is to find those itemsets whose occurrences exceed a predefined threshold in the database; those itemsets are called frequent or large itemsets. The second problem is to generate association rules from those large itemsets with the constraints of minimal confidence. Suppose one of the large itemsets is Lk, Lk = {I1, I2, … , Ik}, association rules with this itemsets are generated in the following way: the first rule is {I1, I2, … , Ik-1} {Ik}, by checking the confidence this rule can be determined as interesting or not. Then other rule are generated by deleting the last items in the antecedent and inserting it to the consequent, further the confidences of the new rules are checked to determine the interestingness of them. Those processes iterated until the antecedent becomes empty. In recent years, XML has become popular. With XML becoming a standard to represent semi-structured data and the increasing amount of XML data available, XML data become an important data mining domain [2], and it is of interest to develop association rule mining methods for such data. Knowledge is extracted out of this XML data and this has to be integrated so as to produce interesting global solutions.

2. Related Work

Association rule mining (ARM) discovers associations between items [1, 3]. Given two distinct sets of items, *X* and *Y*, we say *Y* is associated with *X* if the appearance of *X* implies the appearance of *Y* in the same context. ARM outputs a list of association rules of the format *XÞY*, where *XÞY* has a predetermined support and confidence. Many ARM algorithms are based on the well-known Apriori [1] algorithm. In Apriori, rules are generated from itemsets, which in turn are formed by grouping items that co-occur in instances of data. The prototypical application of ARM is market basket analysis in which items that are frequently purchased together are discovered in order to aid grocers in layout of items. The problem of association rule mining on XML data has been addressed in [4, 5]. However, the rules to be mined are specific ones with antecedent and consequence limited to particular elements.

Manuscript received May 5, 2008
Manuscript revised May 20, 2008

XML data exchange among applications means decentralization of information and as a consequence the distribution aspect of XML data is destined to play an important role. Application such as Web services, e-commerce applications, or the management of large-scale directories are nowadays deployed on the Internet. In addition, the need of interoperability among distributed applications makes XML a good candidate acting as a universal language that every system can comprehend. Nevertheless, distributing data over several sites implies many challenges [6] including heterogeneity, openness, security, scalability, failure handling, concurrency, and transparency. In the emerging networked knowledge environment, the relevant data for many computations may reside on a number of geographically distributed databases that are connected by communication networks. Kargupta et al. [7] proposed the Collective Data Mining (CDM) framework to learn from heterogeneous data sites. The goal of the CDM framework is to generate accurate global models, which one would get if data were centralized/ combined, in a distributed fashion. Chan and Stolfo [8] presented meta-learning techniques, which try to learn a global classifier based on local models built from local data sources, for mining homogeneous distributed datasets. A global computation has to be made from these distributed workstations and the relevant knowledge is obtained for decision making.

3. Parallel and Distributed Data Mining

Parallel and distributed data mining is the process of analyzing the data in distributed workstations, and finding useful and novel knowledge, which is highly supported by most of data warehouses or individual warehouse. If we combine all the data in different databases into a single one and mine the large single data warehouse, then it may hide some features in a database and lose some useful pattern. The huge dataset after integrating will be difficult to deal with and its data may not be stored into memory at a time. To mine large and distributed data sets, it is important to investigate efficient distributed algorithms to reduce the communication overhead, central storage requirements, and computation times. Parallel and Distributed data mining is the mining of distributed data in a parallel environment [11]. It intends to obtain global knowledge from local data at distributed sites. Researchers expect parallelism to relieve current Association Rule Mining (ARM) methods from the sequential bottleneck, providing scalability to massive data sets and improving response time. Achieving good performance on today's multiprocessor systems is not trivial. The main challenges include synchronization and communication minimization, workload balancing, finding good data layout and data decomposition, and disk I/O minimization. Many parallel data mining algorithms inherits this property from Apriori, which is why most parallel data mining algorithms are said to be a variation of Apriori [12]. Writing parallel data mining algorithms are a non-trivial task. The main challenges associated with parallel data mining include

i. Minimizing I/O

ii. Minimizing synchronization and communication

Effective load balancing In a distributed environment, data sites may be homogeneous, i.e., different sites containing data for exactly the same set of features, or heterogeneous, i.e. different sites storing data for different set of features, possibly with some common features among sites. Association rules are used for the mining process with XML dataset in distributed locations. The interesting patterns are identified locally which sometimes may be uninteresting globally. An integration of the obtained knowledge is made so that interesting patterns are generated for making useful decisions.

4. Association rules in XML Data

Association rule mining was mainly used for market basket analysis. The problem of mining association rules can be explained as follows: There is the itemset I = i1, i2,. .., in, where I is a set of n distinct items, and a set of transactions D, where each transaction T is a set of items such that T _ I. Table 1 gives an example where a database D contains a set of transactions T, and each transaction consist of one or more items.

Table 1: An Example Database

tid	**items**
1	{bread, butter, milk}
2	{bread, butter, milk, ice cream}
3	{ice cream, coke}
4	{battery, bread, butter, milk}
5	{bread, butter, milk}
6	{battery, ice cream, bread, butter}

An association rule is an implication of the form X ⇨ Y, where X, Y C I and X Π Y = . The rule X ⇨ Y has support s in the transaction set D if s% of transactions in D contain X [Y. The support for a rule is defined as support(X[Y). The rule X) Y holds in the transaction set D with confidence c if c% of transactions in D that contain X also contain Y. The confidence for a rule is defined as support(X[Y)/support(X).

The association rule from XML data with a sample XML document is considered [10]. We refer to the sample XML document, depicted in Figure 1 where information about

the items purchased in each transaction are represented. For example, the set of transactions are identified by the tag <transactions> and each transaction in the transactions set is identified by the tag <transaction>. The set of items in each transaction Figure 1: Transaction document (transactions.xml) are identified by the tag <items> and an item is identified by the tag <item>. Consider the problem of mining all association rules among items that appear in the transactions document as shown in Figure 2. With the understanding of traditional association rule mining we expect to obtain the large itemsets document and association rules document from the source document.

Let the minimum support (minsup) = 30% and minimum confidence (minconf) = 100%.

```
<transactions>
        <transaction id=1>
                <items>
                        <item> i1</item>
                        <item> i4</item>
                        <item> i7</item>
                </items>
        </transaction>

        <transaction id=2>
                <items>
                        <item> i2</item>
                        <item> i3</item>
                        <item> i5</item>
                </items>
        </transaction>
        <transaction id=3>
                <items>
                        <item> i1</item>
                        <item> i3</item>
                        <item> i7</item>
                </items>
        </transaction>
        <transaction id=4>
                <items>
                        <item> i2</item>
                        <item> i5</item>
                </items>
        </transaction>
        <transaction id=5>
                <items>
                        <item> i1</item>
                        <item> i5</item>
</items>
        </transaction>
```

Figure 1. Sample XML data

The association rules generated by the queries are shown in Figure 2.

```
<rules>
        <rule support="0.3" confidence="1.0">
                <antecedent>
                        <item> d </item>
                        <item> b </item>
                </antecedent>
                <consequent>
                        <item> c </item>
                </consequent>
        </rule>
        <rule support="0.3" confidence="1.0">
                <antecedent>
                        <item> d </item>
                        <item> c </item>
                </antecedent>
                <consequent>
                        <item> b </item>
                </consequent>
        </rule>
        <rule support="0.3" confidence="1.0">
                <antecedent>
                        <item> b </item>
                        <item> c </item>
                </antecedent>
                <consequent>
                        <item> d </item>
                </consequent>
        </rule>
</rules>
```

Figure 2. Association Rules document

The data inside the rules document is self describing. For example, the set of rules are identified by the tag <rules> and each rule is identified by the tag <rule> with two attributes support and confidence to describe the strength of the rule. Inside the tag <rule>, there are two sub-tags <antecedent> and <consequent> which are used to identify the items, antecedent or consequent of the rule.

5. Knowledge Integration with XML Data

The knowledge discovery process takes the raw results from data mining (the process of extracting trends or patterns from data) and carefully and accurately transforms them into useful and understandable information. The term Knowledge Data Discovery (KDD) is increasingly being used as a synonym for data mining. It is a more descriptive term and can be applied to all the activities and processes related the discovering of useful

knowledge from the data. Using a combination of techniques - including statistic analysis, neuronal logic, diffuse logic, multidimensional analysis, data visualization and intelligent agents- the KDD can discover useful patterns to develop models that can predict behaviors or consequences, in a large variety of knowledge spheres. The mined outcome that is obtained from the distributed workstations in the form of knowledge for XML data is integrated based on the following schema.

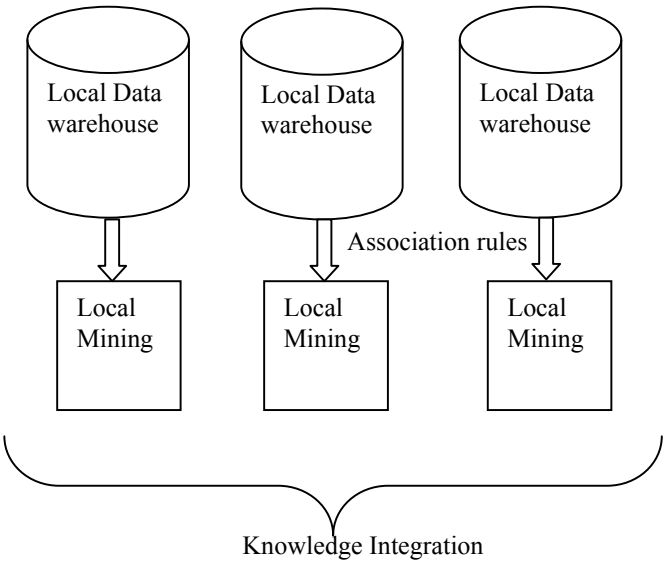

Figure 3. Schematic representation of knowledge integration

Let $D_1,D_2,\ldots D_n$ be n databases in the n branches $B_1,B_2,\ldots B_n$ of a company, respectively; and LI be the set of local patterns (local instances) from $D_i(i=1,2,\ldots n\_)$. The global exceptional patterns of interest in the local patterns are identified. Let $LP_1,LP_2,\ldots LP_n$ be the corresponding local patterns which are mined from every database. And minsupp be the user specified minimal support in the database D_i (i=1,2,…n). For each pattern P, its support in D_i is denoted by $Supp_i(P)$. The average vote of local patterns is given by

$$Aver\ votes = \frac{\sum_{i=1}^{Num(GP)} Num(Pi)}{Num\ (GP)}$$

GP means the Global Patterns, set of all patterns from each data warehouse ie GP= { $LP_1U\ LP_2U\ldots ULP_n$} and Num(GP) is the number of patterns in GP [9]. The global support of a pattern is given by

$$SuppG(P) = \frac{\frac{\sum_{i=1}^{Num(P)} Suppi\ (P) - minsupt}{1 - minsupt}}{Num(P)}$$

SuppG(P) is the global support of a pattern. This gives a method to compute the global patterns from the locally generated knowledge. The global knowledge could be based on the interestingness measures. Exceptional patterns reflect the individuality of branches within an interstate company. An algorithm is used to search all the significant exceptional patterns from the given n local patterns.

Step :1 Generates the set of patterns from each database.

Step :2 Counts each pattern's votes and the average votes of patterns

Step :3 Generates the candidate exceptional patterns

Step :4 Calculate all the candidate exceptional patterns by their SuppG(P) values.

Step :5 Rank the candidate exceptional patterns by their SuppG(P) values.

Step :6 Output all the exceptional pattern which satisfy the users requirement and have high rank.

Based on this algorithm the patterns are taken and globalized integrated result is obtained by eliminating the uninteresting ones.

6. Performance Evaluation

Data mining will be essential for discovering new knowledge from many XML resources likely to arise in the next few years. Given the irony that humans produce far more data than they can ever analyse alone, the development of XML mining techniques must keep pace with development and implementation of XML itself. Experimental analysis is made using association rules with XML dataset on distributed locations in a parallel environment.

An analysis is made with the number of processors involved and the time taken. The communication delays for the distributed workstations are also considered. Additionally, when individual processors were loaded heavily by other running processes in background, the processing time was increased, but the communication delay was observed not to increase in a proportional manner. This can be because decrease in processing capability increases the writing and reading time and not

the communication delay. Moreover communication is done once, whereas the decrease in processing speed of a processor effects processing time of all the passes on the slow processor. Also in an heterogeneous environment the fast processors can still communicate at fast speeds. Figure 4 demonstrates this effect. As one of the processor was loaded more and more the execution time of the system increases but the communication delay shows minor variation.

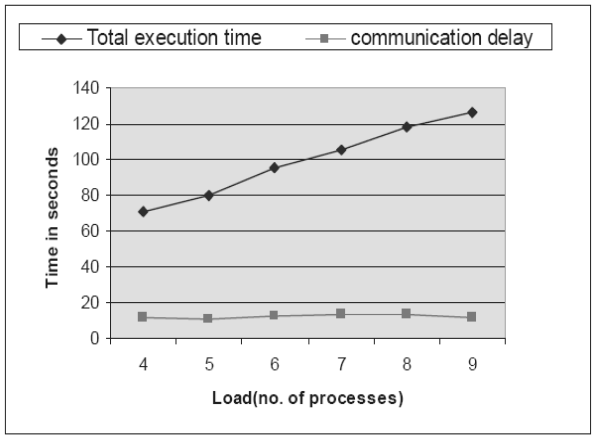

Figure 4. Load Vs. Time

Experiments were also performed by increasing the minimum support and the time taken is found. It can be observed that the performance of these algorithms largely depends on the number of frequent itemsets. For lower values of minimum support, it is expected to have many frequent itemsets, and this number will decrease as the minimum support increases. So the running time decreases as the minimum support increases.

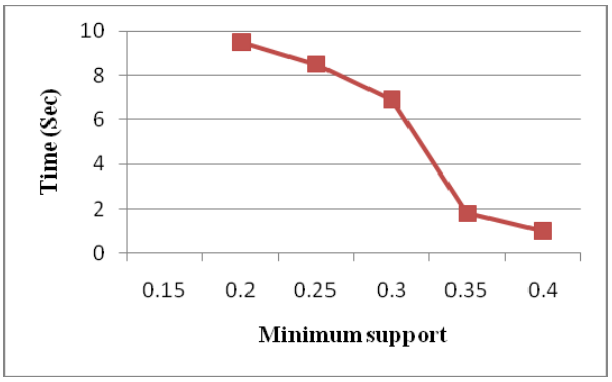

Figure 5. Minimum support Vs. Time

In XML data the mining is carried out. By keeping the minimum support as a constant to 0.3 the confidence is increased and the number of itemsets generated is analysed as shown in Figure 6.

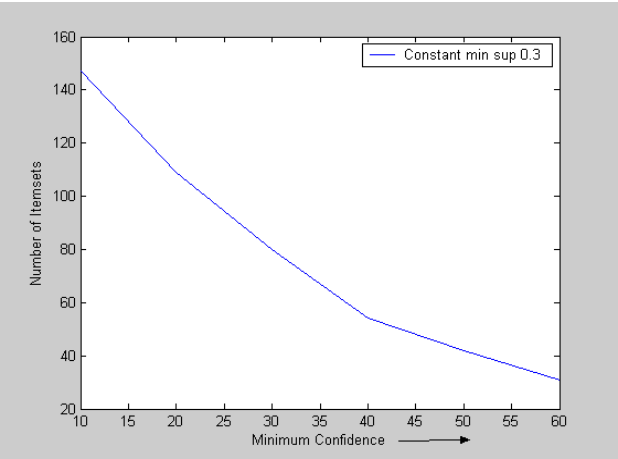

Figure 6. Minimum confidence Vs. Itemsets

The knowledge is thus integrated from the distributed workstations and then sends back to the client for the calculation of response time. The time taken for integration, time taken for mining and the time taken to allocate data will be the different parameters taken for the calculation of response time.

7. Conclusion

The main motivation for eXtensible Markup Language (XML) is it provides a markup language more conducive to data exchange. As semi-structured data, it is self-describing. An XML document has document type definitions (DTD) that define the structure of the document and what tags might be used to encode the document. Due to this advantage, XML will be the dominant format in a few years. This paper hence has done the mining process in a parallel and distributed environment using association rules with this XML data. The knowledge that is obtained out of this mining process is then integrated together by finding out the exceptional patterns, identifying the interesting patterns locally and then global reduction is carried out.

References

[1] Agrawal, R., Imielinski, T., and Swami, A. N. 1993. Mining association rules between sets of items in large databases. In Proceedings of the 1993 ACM SIGMOD International Conference on Management of Data, 207-216.

[2] A. Termier, M-C. Rousset, M. Sebag, “TreeFinder: a First Step towards XML Data Mining”, in Proceedings of IEEE International Conference on Data Mining, 2002.

[3] Agrawal R., Mannila H., Srikant R., Toivonen H., and Inkeri Verkamo A. Fast discovery of association rules. In *Advances in Knowledge Discovery and Data Mining*, pages 307-328. AAAI/MIT Press, 1996.

[4] D. Braga, A. Campi, S. Ceri, M. Klemettinen, PL. Lanzi, "Mining Association Rules from XML Data", in Proceedings of DEXA 2002 (DaWaK), LNCS 2454, Aixen- Provence, France, Sep. 2002, pp. 21-30.

[5] D. Braga, A. Campi, S. Ceri, M. Klemettinen, PL. Lanzi, "A Tool for Extracting XML Association Rules from XML Documents", in Proceedings of IEEE-ICTAI 2002, Washington DC, USA, Nov. 2002, pp. 57-64.

[6] G. Coulouris, J. Dollimore, T. Kindberg Distributed Systems: Concepts and Design., Third Edition, Addison-Wesley.

[7] Kargupta, H., Park, B., Hershbereger, D., & Johnson, E. (2000). Collective data mining: A new perspective toward distributed data mining. In H. Kargupta and P. Chan (Eds.), *Advances in distributed data mining*, 133–184. AAAI/MIT.

[8] Chan, P. C., & Stolfo, S. (1993). Meta-learning for multistrategy and parallel learning. *Proceedings of the Second International Workshop on Multistrategy Learning*.

[9]. Chengqi Zhang , Meiling Liu, Wenlong Nie, and Shichao Zhang, Identifying Global Exceptional Patterns in Multi-database Mining, IEEE computational Intelligence.

[10] Mohammed J. Zaki, Charu C. Aggarwa, "XRules: An Effective Structural Classifier for XML Data", Rensselaer Polytechnic Institute.

[11] S. Krishnaswamy, S.W. Loke2, A. Zaslavsky," Cost Models for Distributed Data Mining", School of Computer Science & Software Engineering, Monash University, 2004.

[12] Chao Wang and Srinivasan Parthasarathy, "Parallel Algorithms for Mining Frequent Structural Motifs in Scientific Data", The Ohio State University, 2004.

[13] Mohammed J.Zaki, "Parallel and Distributed Association Mining: A Survey", IEEE Concurrency, 2000.

Sujni Paul obtained her Bachelors degree in Physics from Manonmanium Sundaranar University during 1997 and Masters Degree in Computer Applications from Bharathiar University during 2000. She is currently pursuing her PhD in Data Mining in the Department of Computer Applications, Karunya University, Coimbatore, India. She is working in the area of parallel and distributed data mining. She was working in Karunya University as Senior Lecturer for 4.5 years and a research scholar at present in the same University.

Dr. V Saravanan obtained his Bachelors degree in Mathematics from University of Madras during 1996 and Masters Degree in Computer Applications from Bharathiar University during 1999. He has completed his PhD in Computer Science in the Department of Computer Science and Engineering, Bharathiar University during 2004. He specialized on automated and unified data mining using intelligent agents. His research area includes data warehousing and mining, software agents and cognitive systems. He has presented many research papers in National, International conferences and Journals and also guiding 3 researchers leading to their PhD degree. He has totally 8 years experience in teaching including 3 years as researcher in Bharathiar University. He is the member of Computer Society of India, Indian Association of Research in Computing Sciences and many professional bodies. At present, he is working as Associate Professor & HOD of the Department of Computer Applications of Karunya School of Computer Science and Technology in Karunya University.