Integration of Data Mining Systems using Sequence Process

Ram Prasad Chakraborty,

Assistant Professor, Computer Science & Engineering, Amrapali Group of Institutes, Haldwani, India.

Abstract- In this paper we discuss about the large database has been a major concern in research community, due to difficulty of analyzing huge numbers of data using only OLAP tools. This process implies lots of computational power, memory space, which can only be provided by parallel computers. Data mining consists of finding interesting trends or patterns in large datasets. We present a discussion of how database technology can be integrated to data mining techniques. We also discuss several advantages of addressing data consuming activities through the patterns of parallel database server and data mining technology.

Keywords- Data Mining, DBMS, KDD, OLAP.

1. INTRODUCTION

Data mining consists of finding interesting trends or patterns in large datasets, in order to guide decisions about future activities. There is a general expectation that data mining tools should be able to identify these patterns in the data with minimal user input? The patterns identified by such tools can give a data analyst useful and unexpected insights that can be more carefully investigated subsequently, perhaps using other decision support tools. Data mining technique have increasingly been studies. Especially in their application in real world database. One typical problem is that databases tend to be very large, and these techniques often repeatedly scan the entire set. Sampling has been used for a long time, but subtle differences among sets of objects become less evident. This work provides an overview of some important data mining techniques and their applicability on large databases. This approach has been a major concern of several researches, because it represents a very natural solution since DBMSs have been successfully used in business management and currently may store valuable hidden knowledge. It is also closely related to the subareas of artificial intelligence called *knowledge discovery* and *machine learning*. The important distinguishing characteristic of data mining is that the volume of data is very large; although ideas from these related areas of study are applicable to data mining problems, *scalability with respect to data size* is an important new criterion. An algorithm is scalable if the running time grows (linearly) in proportion to the dataset size, given the available system resources (e.g., amount of main memory and disk). Old algorithms must be adapted or new algorithms must be developed to ensure scalability. It makes the use of parallelism even more relevant to provide a way of processing long running tasks in a timely manner. In this context, parallel database systems come to play an important role, because they can over, among other advantages, transparent and painless implementation of parallelism to process large data sets. It is important to notice that, when we mention the use of large amounts of information in data mining, we are not referring to usual large DBMSs, which can reach more than one terabyte of data. As data mining methods often repeatedly scan the data set, mining in such a large database is not cited in the literature yet.

2. DATA MINING INTEGRATION

In the real world, data mining is much more than simply applying to one of the algorithms. Data is often noisy or incomplete, and unless this is understood and corrected for, it is likely that many interesting patterns will be missed and the reliability of detected patterns will be low. Further, the analyst must decide what kinds of mining algorithms are called for, apply them to a well-chosen subset of data samples and variables (i.e., tuples and attributes), digest the results, apply other decision support and mining tools, and iterate the process. The knowledge discovery process, or short KDD process, can roughly be separated into four steps. The raw data first undergoes a data selection step, in which we identify the target dataset and relevant attributes. Then in a data cleaning step, we remove noise and outliers, transform field values to common units, generate new fields through combination of existing fields, and bring the data into the relational schema that is used as input to the data mining activity. The data cleaning step might also involve a de- normalization of the underlying relations. In the data mining step, we extract the actual patterns. In the final step, the evaluation step, we present the patterns in an understandable form to the end user, for example through visualization. The results of any step in the KDD process might lead us back to an earlier step in order to redo the process with the new knowledge gained. We present four classes of data mining techniques typically used in a variety of well-known applications and researches currently cited in the database mining community. They certainly do not represent all mining methods, but are a considerable portion of them when a large amount of data is considered.

2.1 Based on Classification Process

Abstractly, the classification problem is this: Given that items belong to one of several classes, and given past instances (called training instances) of items along with the classes to which they belong, the problem is to predict the class to which a new item belongs. The class of the new instance is not known, so other attributes of the instance must be used to predict the class. The rules may be the following forms:

$\forall$ *person P, P.degree= masters* and *P.income*>75, 000$\Rightarrow$*P.credit= excellent*

$\forall$*person P, P.degree= bachelors* or

(*P.income* $\geq$ 25, 000 and *P.income* $\leq$ 75, 000) $\Rightarrow$*P.credit= good.*

Classification algorithm follows two methods that are widely used for data mining technique: Decision tree and Neural networks. Classification example: below in fig. 1.

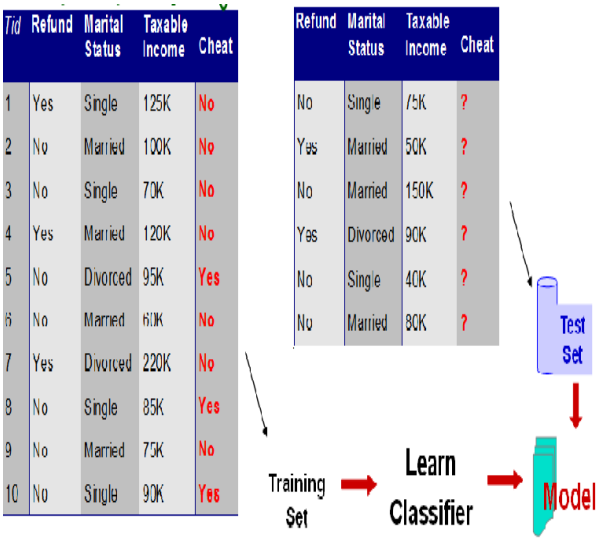

Fig-1:Classification Sample

2.2 Based on Decision Trees

The type of rules that we discuss can be represented by a tree, and typically the tree itself is the output of the data mining activity. Trees that represent classification rules are called classification trees or decision trees and trees that

represent regression rules that are called regression trees. Decision tree methods are a kind of machine learning algorithm that uses a divide-and-conquer approach to classify cases using a tree-based representation

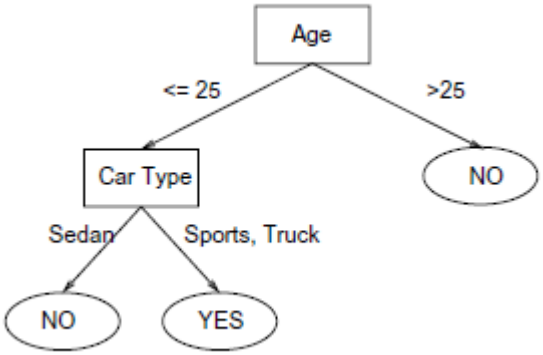

Fig-2: Insurance Risk Example Decision Tree

As an example, consider the decision tree shown in Fig. 2. Each path from the root node to a leaf node represents one classification rule. For example, the path from the root to the leftmost leaf node represents the classification rule: If a person is 25 years or younger and drives a sedan, then he is likely to have a low insurance risk." The path from the root to the right-most leaf node represents the classification rule: If a person is older than 25 years, then he is likely to have a low insurance risk."Tree-structured rules are very popular since they are easy to interpret. There exits efficient algorithms to construct tree structured rules from large database.

2.3 Neural Networks

This method based on artificial neural networks that provide a general and practical method for learning functions, which are represented by continuous attributes, discrete or vectors. Basically, neural networks have been used to interpret visual scenes, voice recognition, and they are not only used for classification. Neural network describe the difficult interpretation, tends to over fit the data, extensive amount of training data and lot of data preparation and also work with all data types that are shown in fig.3.

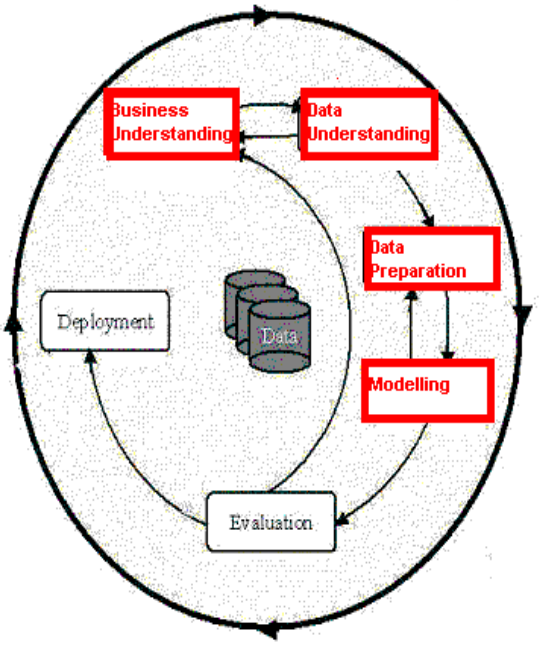

Fig-3: Data mining evaluation model

Neural networks typically require more time to finish than, say, decision trees algorithms. Training times vary depending on the number of training cases, the number of weights in the network, and the settings of many learning algorithm parameters. Finally the ability to understand learned target function is not important. Learned neural networks are less easily communicated to humans than learned rules.

3. ASSOCIATION RULES

Association rules can be used in different ways likes when a customer buys a particular book, an online shop may suggest associated books. A grocery shop may decide to place bread close to milk, since they are often bought together, to help

shoppers finish their task faster. Or the shop may place them at opposite ends of a row, and place other associated items in between to tempt people to buy those items as well, as the shoppers walk from one end of the row to the other. A shop that offers discounts on one associated item may not offer a discount on the other, since the customer will probably buy the other anyway.

Rules:

Support: "is a measure of what fraction of the population satisfies both the antecedent and the consequent of the rule"

Example:

People who buy hotdog buns also buy hotdog sausages in 99% of cases. = High Support

People who buy hotdog buns buy hangers in 0.005% of cases. = Low support.

Situations where there is high support for the antecedent are worth careful attention E.g. Hotdog sausages should be placed in near hotdog buns in supermarkets if there is also high confidence.

Confidence: "is a measure of how often the consequent is true when the antecedent is true." Example: 90% of Hotdog bun purchases are accompanied by hotdog sausages. High confidence is meaningful as we can derive rules. Hotdog bun ➔ Hotdog sausage 2 rules may have different confidence levels and have the same support. E.g. Hotdog sausage ➔ Hotdog bun may have a much lower confidence than Hotdog bun ➔Hotdog sausage yet they both can have the same support.

3.1 Clustering

Given a set of data points, each having a set of attributes, and a similarity measure among them, find clusters such that :

Data points in one cluster are more similar to one another.

Data points in separate clusters are less similar to one another.

Similarity Measures: Euclidean Distance if attributes are continuous. Other Problem-specific Measures. Fig. 4. describe the clustering process with 3D spacing.

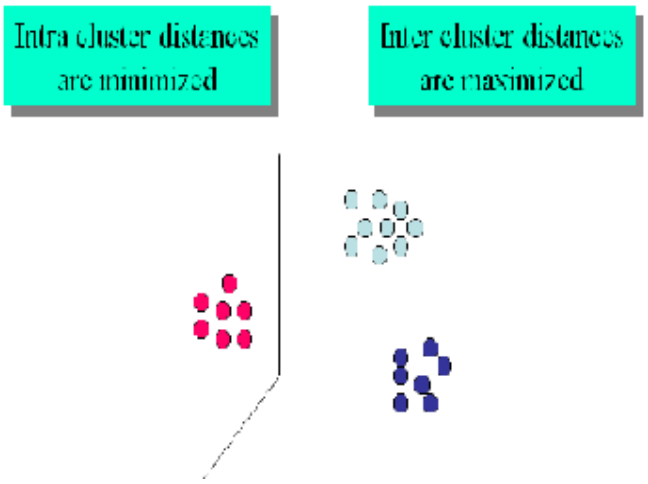

Fig-4: Clustering in 3D space

Clustering algorithms find groups of items that are similar. It divides a data set so that records with similar content are in the same group, and groups are as different as possible from each other. Example: Insurance company could use clustering to group clients by their age, location and types of insurance purchased.

3.1 Sequential Patterns

Given is a set of *objects*, with each object associated with its own *timeline of events*, find rules that predict strong sequential dependencies among different events and the discovery of sequential patterns has been motivated by applications in retailing industry, including attached mailing and add-on sales, and in the medical domain likes:

3.1.1 In telecommunications alarm logs

Inverter-Problem➔Excessive-Line-Current (Rectifier-Alarm) ➔ (Fire-Alarm)

3.1.2 In point-of-sale transaction sequences

Computer Bookstore:

(Intro-To-Visual-C) (C++_Primer) ➔ (Perl-for-dummies, Tcl-Tk)

Athletic Apparel Store:

(Shoes) (Racket, Racket ball) ➔ (Sports-Jacket)

4. RESEARCH BASED ARCHITECHTURE

4.1 System and Data Mining

The architecture of a database system is greatly influenced by the underlying computer system on which it runs, in particular by such aspects of computer architecture as networking, parallelism, and distribution. Database technology has been successfully used in traditional business data processing. Companies have been gathering a large amount of data, using a DBMS system to manage it. Therefore, it is desirable that we have an easy and painless use of database technology within other areas, such as data mining. DBMS technology offers many features that make it valuable when implementing data mining applications. For example, it is possible to work with data sets that are considerably larger than main memory, since the database itself is responsible for handling information, paging and swapping when necessary That describe in fig. 5.

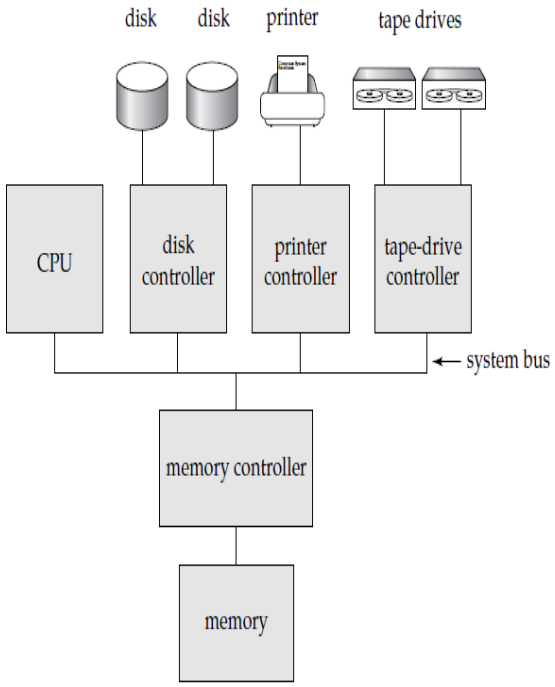

Fig-5: Centralized system

5. DATA MINING AND PARALLEL DATABASE SYSTEM

The transaction requirements of organizations have grown with increasing use of computers. Moreover, the growth of the World Wide Web has created many sites with millions of viewers, and the increasing amounts of data collected from these viewers has produced extremely large databases at many companies. Given that it is desirable that a data mining process handles a large volume of data, parallel algorithms are needed to provide scalability in order to end the process in a timely manner. There are a variety of data mining algorithms constructed to run-in parallel, taking advantage of parallel architectures using specific programming routines. Alternatively, parallel database systems can provide parallelization in a transparent, painless manner to the application.

A. Some advantages of using parallel database system:-

1 The implementation becomes simpler:

There is no need to use parallel routines such as MPI libraries. The parallel DBMS is responsible itself for parallelizing queries that are issued against it. We have to structure our queries so that they can fully exploit parallelism offered by a particular DBMS.

2 Opportunity for database fragmentation:

Database fragmentation provides many advantages related to the reduced number of page accesses necessary for reading data. Irrelevant data could be just ignored depending on the alter specified by SQL statements. Furthermore, some DBMSs can process each partition in parallel, using different execution plans when applicable. In general, DBMS systems provide automatic management of such functionality.

B. Some disadvantages of using parallel database system

1. Less control of how parallelization will occur-

Although there is the possibility of some customizations, such as setting the nodes that will participate in the process, the DBMS itself is in charge for parallelization.

2. Overhead of the database system kernel-

A kernel of a database system is designed to handle a large set of operations, such as OLTP transactions, OLAP queries, etc. Although it is possible to minimize the overhead and customize the environment to take the best available advantages of the corresponding architecture, there will always exist some functionality implemented that is not applicable to the data mining algorithm, which can degrade performance when compared to access to flat files.

6. DATA MINING FUTURE STUDYS

In this paper describe about the problem of discovering patterns from a database. There are several other equally important data mining tasks, some of which we discuss briefly below. Two example data mining products IBM Intelligent Miner and Silicon Graphics Mine set: Both products over a wide range of data mining algorithms including association rules, regression, classification, and clustering. The emphasis of Intelligent Miner is on scalability the product contains versions of all algorithms for parallel computers and is tightly integrated with IBM's DB2 database system. Mine set supports extensive visualization of all data mining results, utilizing the powerful graphics features of SGI workstations.

Dataset and feature selection: It is often important to select the `right' dataset to mine. Dataset selection is the process of finding which datasets to mine. Feature selection is the process of deciding which attributes to include in the mining process.

Sampling: One way to explore a large dataset is to obtain one or more samples and to analyze the samples. The advantage of sampling is that we can carry out detailed analysis on a sample that would be infeasible on the entire dataset, for very large datasets. The disadvantage of sampling is that obtaining a representative sample for a given task is difficult; we might miss important trends or patterns because they are not reflected in the sample. Current database systems also provide poor support for efficiently obtaining samples. Improving database support for obtaining samples with various desirable statistical properties is relatively straightforward and is likely to be

available in future DBMSs. Applying sampling for data mining is an area for further research.

Visualization: Visualization techniques can significantly assist in understanding complex datasets and detecting interesting patterns, and the importance of visualization in data mining is widely recognized.

CONCLUSIONS

Data mining consists of finding interesting trends or patterns in large datasets, in order to guide decisions about future activities. Data mining and its application on large databases have been extensively studied due to the increasing difficulty of analyzing large volumes of data using only OLAP tools. This difficulty pointed out the need of an automated process to discover interesting and hidden patterns in real-world data sets. The ability to handle large amounts of information has been a major concern in many recent data mining applications. Parallel processing comes to play an important role in this context, once only parallel machines can provide sufficient computational power, memory and disk I/O. We described some important data mining techniques, presenting brief descriptions about them and showing how each one can contribute to the pattern discovery process. Furthermore, we presented several advantages of implementing a data mining method using a DBMS instead of conventional flat files. Our practical work exploited many specific characteristics of DBMSs, providing a tightly-coupled integration of a data mining technique with a parallel database server using a complex application.

REFERENCES

[1] Agrawal, R., &Srikant, R., Fast Algorithms for mining association rules, Proc. of the 20th VLDB Int. Conf., Santiago, Chile, 1994.

[2] Chen, M. S., Han, J., & Yu, P. S., Data Mining: An Overview from Database Perspective, IEEE Trans. on Knowledge and Data Engineering, Vol. 8, No. 6, pp. 866-883, 1996

[3] Agrawal, R., Imielinski, T., & Swami, A., Mining association rules between sets of items in large databases, Proc. of Int. Conf. ACM SIGMOD, Washington D. C. pp. 207-216, 1993.

[4] Agrawal, R., Gehrke, J., Gunopulos, & D., Raghavan, P., Automatic Subspace Clustering of High Dimensional Data for Data Mining Applications, Proc. of the ACM SIGMOD Int. Conf. on Management of Data, Seattle, Washington, 1998.

[5] Agrawal, R., Metha, M., Shafer, J., &Srikant, R., The Quest Data Mining System, Proc. of the 2nd Int. Conf. on Knowledge Discovery in Databases and Data Mining, Portland, Oregon, 1996.

[6] Hallmark, G., Oracle Parallel Warehouse Server, Proc. of ICDE, pp. 314-320, 1997.

[7] Han, J., Fu, Y., Koperski, K., Melli, G., Wang, W., &Zaane, O., Knowledge Mining in Databases: An Integration of Machine Learning Methodologies with Database Technologies, Canadian AI Magazine, 1995.

[8] Han, J., Fu, Y., Wang, W., Chiang, J., Gong, W., Koperski, K., Li, D., Lu, Y., Rajan, A., Stefanovic, N., Xia, B., &Zaiane, O., DBMiner: A System for Mining Knowledge in Large Relational Databases. Proc. Int. Conf. on KDD, Portland, Oregon, 1996.

[9]Silberschatz, Korth, Sudarshan, "Database System Concepts", 5th Edition, McGraw Hill, 2005

[10]http://www.twocrows.com/glossary.htm, "Two Crows, Data Mining Glossary"

[11]http://en.wikipedia.org/wiki/Data_mining, "Wikipedia"

[12]http://phoenix.phys.clemson.edu/tutorials/excel/regression.html.

[13]http://wwwmaths.anu.edu.au/~steve/pdcn.pdf.

[14]Silberschatz–Korth–Sudarshan Database System Concepts, Fourth Edition

BIOGRAPHIES

Ram Prasad Chakraborty, is M-Tech (S/W Engg.) From SRM University, Chennai, now he is working as an Assistant Professor at Amrapali Institute wani (U.K). He is also a working as a Motivational speaker & Social worker.
Email:-durgapurblog@gmail.com
Mobile:-09758048641.

