ISSN:0975-9646

Sales Prediction Using Effective Mining Techniques

Nirav Shah[1], Mayank Solanki[2] , Aditya Tambe[3] , Dnyaneshwar Dhangar[4]

[#123] Computer Engineering Department
[#4] Professor Computer Engineering Department
Rajiv Gandhi Institute of Technology
Andheri (west),Mumbai,India

***Abstract*— Data mining is extraction of hidden and predictive information from huge database; it is a strong new technology with great potential to help companies to focus on the most important information in their data warehouses. It captures the browsing behavior of users at a Company. So, proposed system is used to find most frequent combinations of item present in company. This will help in marketing and sales. This system can be used to discover interesting cross-sells and related products. The proposed system uses apriori algorithm with modification which will make algorithm more efficient. The analyst can perform data mining and extraction and finally conclude the result and make appropriate decision for company.**

***Keywords*— Association rules, Apriori algorithm, Distributed data mining**

I. INTRODUCTION

Mining frequent item sets from the large transactional database is a very critical and important task. Applications requiring large amount of data processing[1], consists of two huge problems, one is high storage and its management and the other one is the processing time, due to increase in data. Distributed databases do the work of solving the first problem to a tremendous extent but second problem boosts. As current era is of communication and association and people are interested in storing large data on networks, and hence, researchers are introducing various algorithms to boost the throughput of output data over distributed databases. In our research, we are introducing a algorithm to practice large amount of data at the various servers of same company that lie at different locations and collecting the practiced data on main server machine as much as admin is requiring. The local copy of found data is provided to the users if he/she needs it again, this allows causing a proxy server where constantly searched items can be saved with the density of their access. This not only grants affording fast access to the data but will also afford to maintain list of recurrently accessed data.

There are several approaches for accessing the data from the various servers, such as direct networked access, mobile agents, client-server techniques and LAN etc. We have used multi-threaded environment to calculate various distributed servers to gather data. For processing of data at the server end, the use of Apriori Algorithm has been done to get the outputs, which are then addressed to the client. At client data from various servers is assembled and then disciplined into data format.

As an association rule mining is defined as the relation between various item sets. Association rule mining takes part in pattern discovery techniques in knowledge discovery and data mining (KDD). As performance of association rule mining is depends upon the frequent item sets mining, thus is necessary to mine frequent item set efficiently. Association rules arrange information of this type in the form of "if-then" statements. These rules are count from the data and, inconsistent the if-then order of logic, association rules are probabilistic simultaneously, the antecedent (the "if" part) and the consequent (the "then" part), an association rule has two numbers that explicit the degree of ambiguity about the rule. In association partition the antecedent and consequent are sets of items (called item sets) that are disjoint (do not have any items in common).An Optimized Distributed Association Rule mining algorithm for geographically distributed data is used in this paper in parallel and distributed environment so that it lowers communication costs.

Association rule mining finds interesting associations relationships among large set of data items. Association rules view aspects value conditions that occur frequently together in a given data set. A classic and extensively-used example of association rule mining is Market Basket Analysis that is main aim and purpose of our system.

II. LITERATURE SURVEY

There is a need to develop a good algorithm which finds the desired information resources and their usage pattern and also to develop a distributed algorithm for geographical data sets that reduces communication cost and communication overhead [3]. There are various algorithms which significantly work in this domain. Some are discussed below.

2.1 *APRIORI ALGOITHM*

Apriori, an association rule mining algorithm innovation has been advanced for rule mining in large firm databases by IBM's Quest project team. An {item set} is a non-empty set of items. They have breakdown the complication of mining association rules into two parts:

1. Find all combinations of items that have transaction support above minimum support. Call those sequence frequent item sets.

2. Use the frequent item sets to achieve the desired rules. The natural concept is that if, say, ABCD and AB are

Frequent item sets, and then we can determine if the Rule AB CD holds by computing the ratio r = support (ABCD)/support (AB).The rule holds only if r >= minimum confidence.

2.2 *CLASSIFICATION*

The process of breaking down a data set into mutually absolute groups such that the representatives of each club are as "close" as possible to one another, and different clubs are as "far" as possible from one another where distance is measured with respect to specific variable(s) you are trying to anticipate for example, a typical classification complication is to break down a database of companies into clubs that are as homogeneous as possible as regards to a creditworthiness variable with values "Good" and "Bad"[3].

2.3 *CLUSTERING LGORITHM*

The process of dividing a data set into mutually exclusive groups such that the members of each group are as "close" as possible to one another, and different groups are as "far" as possible from one another, where distance is consistent with respect to all available variables given databases of sufficient amount and aspect, data mining technology can generate new business opportunities by providing these capabilities.

III. PROPOSED SYSTEM

A typical example of a predictive problem is target marketing. Data mining uses data on past informational expressing to identify the targets most likely to maximize return on investment in future addressing. Unlike other algorithms, O-DAM [Optimized Distributed Association rule Mining] attempt greater performance by minimizing applicant item set. It accomplish this by attract on two major D-ARM [Distributed Association Rule Mining] issues inter communication and integration. Inter Communication is one of the better extensive D-ARM objectives. D-ARM algorithms will achieve better if we can reduce inter communication [4]. D-ARM discovers rules from various geographically assigned data sets. However, the network connection between those data sets isn't as fast as in a parallel environment, so assigned mining usually intent to minimize communication costs. It is designed to operate on files. The algorithm finds the subsets which are common to at least a minimum number C (the cutoff, or confidence threshold) of the item sets.

In proposed system we have assumed 3 branches of a company and manager of that respected branch can have access to its database of that branch to find hidden pattern in database. There is one central server rather than having 3 databases at each branch which stores all records about company and it is managed by admin now interesting question is what access they do? Both manager and admin can alter data i.e. add, delete records in database. Now to find hidden pattern from database, they use system in which have implemented apriori algorithm to find frequent items in database, and they can view the output as report and pie chart. Pie chart is used for better understanding. So manager of each branch can find frequent item in its branch while admin can see or find frequent items in all branches as admin has to take important decisions about company.

We have implemented our system using 2 technologies. One is visual studio 8.0 for front end i.e. to take user input and display output second is Microsoft SQL server, it is used to store data and we have implemented apriori algorithm in database system only using cursor concept. The process is as follows:

1. Person (manager or admin) log in system with user name and password.
2. System will validate user name and password and allows access if they are matched in database.
3. Now person can modify (add, delete, view) data in database.
4. Now to find frequent item it will ask system using report button.
5. Now system will find frequent item in database.
6. Display output as report and pie chart.

3.1 *PURPOSE*

It has become increasingly necessary for users to utilize automated tools in find the desired information assets, and to clue and analyze their usage patterns. Association rule mining is an alive data mining analysis field. However, most ARM algorithms cater to a integrate environment.

Distributed Association Rule Mining (D-ARM) algorithms have been advanced. These algorithms, however, conclude that the databases are either horizontally or vertically divided. In the special case of databases populated from information extracted from extol data; existing D-ARM algorithms cannot discover rules based on higher-order associations between items in divided items that are neither vertically nor horizontally assigned, but rather a hybrid of the two.

In our system admin have access to all branches and thus helps admin to find distributed association rules between the different databases branches located at different locations i.e. Apriori is working in distributed environment.

3.2 *WHAT IS DIFFERENT IN OUR ALGORITHM IMPLEMENTATION?*

Generally the process of extracting association rule mining consists of two parts firstly, mine all frequent item sets pattern each of these pattern should satisfy the minimum support threshold. Once these entire frequent patterns are mined, then only second phase of mining i.e. association rules are produced from these frequent item sets. These association rules must satisfy the minimum support and minimum confidence. This minimum support and confidence should be defined by the user.

A large number of algorithms with different mining efficiencies were proposed by many researchers for generation of frequent item sets. Any algorithm should find the same set of rules though their computational efficiencies and memory requirements may be different. The best known mining algorithm is Apriori algorithm. Apriori algorithm is associated with certain limitations of large database scans. Thus variations of Apriori come into existence.

In our system the item from candidate list immediately removed as soon as it doesn’t satisfy support or confidence condition. Thus, it does not generate unwanted candidate thus takes less time to execute and makes system performance better. Also, rather than scanning entire database user (manger/admin) can specify the beginning and ending date. Only those records who fall between this two dates are only considered for association rule generation, thus make system more efficient. Thus we have added 2 new properties in conventional apriori algorithm.

3.3 *DATA FLOW DIAGRAM*

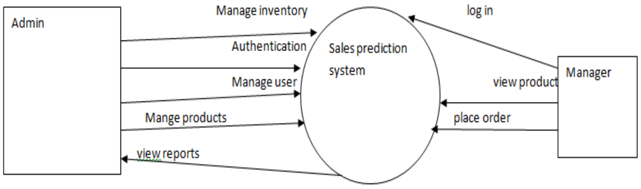

Fig. 1 DFD of our system

A **data flow diagram** (**DFD**) is a graphical representation of the "flow" of data through an information system, modeling its process aspects. A DFD is often used as a preliminary step to create an overview of the system.

Fig 1 shows DFD for our system, it shows what functions of admin and manager are. In our system manager can manage inventory of products present in company of all branches.

IV. CONCLUSION

Sales Prediction Using Effective Mining Techniques is the application of data mining techniques to discover usage patterns from data, in order to understand and better serve the needs of applications. The algorithm used in present research is the Apriori algorithm. This algorithm generates association rules that associate the usage pattern of the clients for a particular data.

The disadvantage of classical apriori is candidate set generation, thus we made 2 changes in algorithm that will make system more efficient.

FUTURE WORK

This system is applicable for very large databases where the available memory space is valuable and requires optimization. It can be further tuned for better performance and efficiency.

It can be applied in applications that deal in mining on live data on daily timely basis such as stock markets, financial statistics collection, weather forecasting etc.

Its application can be utilized for industrial usage where precise pattern study is required with a large data sets to work on and so it can be modified according to their requirements.

REFERENCES

[1] R. Agrawal and R. Srikant , "Fast Algorithms for Mining Association Rules in Large Database", Proc. 20th Int'l Conf. Very Large Databases (VLDB 94), Morgan Kaufmann, 1994,pp. 407-419.

[2] R. Agrawal and J.C. Shafer , "Parallel Mining of Association Rules", IEEE Tran. Knowledge and 16Data Eng. , vol. 8, no. 6, 1996,pp. 962-969;

[3] R. Cooley and B. Mobasher and J. Srivastava, “Web Mining:Information and Pattern Discovery on the World Wide Web”, IEEE International Conference on, pg-558, 1997.

[5] A. Schuster and R. Wolff , "Communication-Efficient Distributed Mining of Association Rules", Proc. ACM SIGMOD Int'l Conf. Management of Data, ACM Press, 2001,pp. 473-484.

[6] R. Agrawal, T. Imielinski, and A. Swami, "Mining Association Rules Between Sets of Items in Large Databases," Proc. ACMSIGMOD Int'l Conf. Management of Data, , May 1993.

[7] M.Z Ashrafi, Monash University ODAM:, “An Optimized Distributed Association Rule Mining Algorithm”, IEEE DISTRIBUTED SYSTEMS ONLINE 1541-4922 © 2004.

[8] Kimball R., Ross M.:, “The Data Warehouse Toolkit, The Complete Guide to Dimensional Modeling”, 2nd edn. John Wiley & Sons, New York (2002)

[9] Ma, Y., Liu, B., Wong, C.K.: Web for Data Mining:, “Organizing and Interpreting the Discovered Rules Using the Web”, SIGKDD Explorations, Vol. 2 (1). ACM Press, (2000) 16-23.

[10] Zaky, M.J., Parthasarathy, S., Ogihara, M., Li, W.:, “New Algorithm for Fast Discovery of Association Rules” Technical Report No. 261, University of Rochester (1997).