RESEARCH ARTICLE OPEN ACCESS

Scope of Efficiency for Association Rule Mining In Cloud Environment

K.Mangayarkkarasi[1], M.Chidambaram[2]
Research Scholar[1], Research and Development Centre,
Bharathiar University, Coimbatore
Assistant Professor[2], Rajah Serfoji College,
Bharathidasan University, Tanjavur
Tamil Nadu-India

ABSTRACT
This paper focuses on one important data mining technique known as association rule mining. With massive amounts of data continuously being collected and stored, many industries are becoming interested in mining association rules from their databases. This paper discusses various association rule mining algorithms that exist today. Based on the factors that affect the efficiency of association rule mining algorithms, this paper derives that cloud technology can be utilized to add more efficiency to association rule mining algorithms.
Keywords:- Association rule mining, cloud computing, ARM algorithms in cloud environment

I. INTRODUCTION

Association rule mining finds interesting association or correlation relationships among a large set of data items. The discovery of interesting association relationships among huge amounts of business transaction records can help in many business decision making processes such as catalog design, store layout etc. Association Rule Mining (ARM) is being given more importance for research because it finds correlation among items in a given data sets and establishes an association between two non overlapping sets of frequently occurring values in a database. Generation of interesting association rules are of use in various domains such as health care, market basket analysis, telecommunication etc.

Now a day's large amount of data is created and generated through real time systems, sensors, sites etc., this trend creates the demand for the advancement in data collection, and storing technology. Hence there is a growing need to run data mining algorithm on very large data sets. To cater to these needs cloud computing seems to be the promising technology. Users can access and deploy cloud applications from anywhere in the world at very competitive costs. Virtualized cloud platforms are often built on top of large data centres. Clouds grow out of the desire to build better data canters through automated resource provisioning. The configuration of this paper is as follows. Section 2 describes the basic concepts of association rule mining. Section 3 discusses the factors those influence the efficiency of the ARM algorithms. Section 4 discusses the concept of parallel and distributed algorithms. Section 5 details the scope of adding efficiency to ARM algorithms through cloud technology and at last concludes.

II. ASSOCIATION RULE MINING

Association rule mining is the discovery of association rules showing attribute-value conditions that occur frequently together in a given set of data. Association analysis is widely used for market basket or transaction data analysis.

The general form of association rules is that A=>B where A,B are set of attribute, value pairs. Here A = $\{a_i\}$ for all i=1,2..n, B = $\{b_j\}$ for all j = 1,2,...m a,b are database tuples such that the tuples in A are likely to be member of B.

The association analysis can be explained with one example. Suppose that the purchase in a mobile shop is under study. If out of 2% of sampled customers purchase a mobile, 60% of them are likely to purchase mobile cover also. This association can be represented as a rule like:

Mobile => Mobile cover [2%, 60%]

[Support=2%, confidence=60%]

This simple association is of single dimension. If one more attribute is added to this transaction then the association rule gets multi dimension. For example say girls of age group 15 to 18 who purchase mobiles are likely to purchase pink colour cover. The association rule that implies this correlation will be as follows.

{age(x,"15...18"),sex(x,"female")}=>
{purchase(x,"pinkcover")

[support=2%, confidence = 60%]

Here x represents a customer of the mobile shop.

Marketing managers are interested to know which items are purchased together. Because by studying the purchase style or pattern they can plan well their item display, store layout etc., so as to improve their business.

There are two major steps in association rule mining. The first step is to find the item sets that occur frequently in transactions .The number of occurrence of the item sets must be greater than or equal to the pre-defined minimum support count. The item sets that satisfy this condition form the frequent item set. The second step is to generate strong association rules. Marketing people are interested to know certain associations. Certain associations hold some purpose or value, such association rules are considered strong association rules. In the second step strong association rules are generated using frequent item sets. The efficiency of association rule generation is majorly decided by the first step of frequent item set generation. Hence most of the algorithms are designed and modified focusing the frequent item set generation.

2.1 Sequential ARM Algorithms

Here we discuss some of the sequential algorithms and their weaknesses. The Apriori algorithm mines simplest form of Boolean association rules from transactional database. Using the Hash Tree data structure it applies bottom up search technique over horizontal data layout. This algorithm iteratively generates all frequent item sets by generating candidate item sets. It needs number of data scans and also generates large number of candidate item sets. Direct Hashing and Pruning algorithm (DHP) [2] Parallel counts 1-itemsets and computes support of 2-itemsets ahead using hash table. It adds k-item sets into C_k only if the k-item sets satisfy minimum threshold and stores them in Hash table. The problem with this algorithm is that only logical pruning can be done. The partition algorithm [3] works in two levels. In level-I it separates the horizontal database into many non-overlapping partitions. Then vertical tid_lists are prepared for each item in each partition. After this all locally frequent item sets are generated via tidlistintersection. Finally all the local frequent item sets are merged together to form the global candidate item set. Level-II identifies the frequent item set by calculating the support count for the global candidate item set generated in level-I This algorithm requires only two scans of the database. But here each partition size is restricted to main memory size.

The SEAR, SPTID, SPEAR and SPINC [4] are based on both Apriori and partition algorithms SEAR uses prefix tree and reduces number of database scans. SPTIP uses prefix tree and straight away executes tid-list intersection. It is too expensive because of join of 2-itemsets which is expensive. SPEAR follows partition algorithm using prefix tree. SPINC uses horizontal data layout and prefix tree to store candidates.

III. FACTORS THAT INFLUENCE THE EFFICIENCY OF ARM ALGORITHMS.

The performance of the ARM algorithms is majorly decided by the factors (1) Time taken to generate frequent item sets.(2) Inter site communication costs(3) Number of scans through the database. Among these factors the first one time taken to generate frequent item sets can be minimized by introducing parallelism to algorithms. Regarding the second factor cost of inter site communication can be reduced through advanced network capabilities and finally to reduce the number of scans through databases there is need for a programming model which suits massive dataset processing.

IV. PARALLEL AND DISTRIBUTED ALGORITHMS

The performance of the sequential algorithms should not degrade with growth in massive datasets. Efficiency increases when communication is minimized and processes are well synchronized. By incorporating data parallelism and task parallelism to ARM algorithms efficiency can be obtained.

4.1. Distributed ARM algorithms

The ARM algorithm in a distributed environment can be represented as parallel processing over partitioned and distributed layout. Consider M as the massive data base to be mined. Let M be split into n partitions and stored in n clusters distributed over a network. Each partition of M be stored in n clusters C_1, C2, Cn as M1, M2…M_n respectively.

Any item set I has the global support count I.Scnt in M and local support count $I.Scnt_k$ in the partition M_k. If T is the given support threshold then when $I.Scnt \geq T$ x number of transactions in M ,I is said to be globally large in M, and when $I.Scnt_k \geq T$ x number of transactions in M_k, I_k is said to be locally large at M_k.

The count distribution (CD) algorithm [5] is an adaptation of the Apriori algorithm in the distributed case. Another interesting distributed ARM algorithm is fast distributed mining of association rules (FDM), [6] that generates a small number of candidate sets and reduces the number of messages to be passed at mining association rules.

V. SCOPE FOR IMPROVISATION TO ARM ALGORITHMS IN CLOUD ENVIRONMENT

Cloud computing provides a virtual platform with elastic resources. It provides hardware, software and datasets dynamically on-demand. In cloud environment programs are sent to where the data is located, rather than copying the data to millions of desktops. The application codes are comparatively much smaller than the datasets they process. Cloud environment avoids large data movements hence enables better network bandwidth utilization. Cloud also relieves massive I/O problem.

Virtualization enables the construction of a more secured, fault tolerant, robust environment than traditional distributed systems. Cloud has several security issues involving assurance and confidentiality of data. Data analysis is being done to extract valuable information from a large volume of data. These analysis techniques are being used by cloud service providers. As mining algorithms require vast amount of data, the single provider architecture suits the purpose of the attackers. The privacy of data in the cloud has become a major concern. But by distributing user data among multiple cloud providers makes data mining a difficult task for the attackers, thereby security can be improved.

The association rules mining from the cloud can be done using sector/sphere framework. Famous IT corporations have their cloud computing architecture. One such example is Google App Engine which is composed of Google File System (GFS), Big table (Hadoop) and Map Reduce.

Map Reduce is a programming model for processing large data sets Map Reduce is a framework for processing the parallelizable problems [7] across huge datasets using a large number of computers. It can take advantage of locality of data, processing data on or near the storage assets to decrease transmission of data. In addition the management of the distributed systems like transferring among nodes, site failures are monitored and maintained by Hadoop, this adds robustness and scalability to the system.

VI. CONCLUSION

Now a day's data are vast and distributed in nature, hence to add efficiency to association rule mining, parallelism and distributed processing need to be introduced. In this paper we reviewed certain sequential and distributed algorithms for association rule mining. The review derived certain factors influencing the performance of ARM algorithms. All those factors can very well be managed in the cloud environment. We gave one sample Google cloud architecture which gives a promising environment to implement efficient parallel ARM algorithms. The implementation of many parallel ARM algorithms in cloud environment need to be studied in future for their efficiency.

REFERENCES

[1] R.Agarwal and J.Shafer, "Parallel mining association rules", IEEE Trans. On knowledge and Data Engg., 8(6):962-969, December 1996, pp. 4-6, 14.

[2] J.S.Park, M.Chen and P.S.Yu, An effective hash based algorithm for mining association rules.In ACM SIGMOD Intl. Conf. Management of Data, May 1995, pp. 176,178-182.

[3] A.Savasere, E.Omiecinski and S.Navathe, "An efficient algorithm for mining association rules in large databases", in the proceeding of 21st VLDB Conference, 1995, pp. 1-5.

[4] A.Mueller, "Fast sequential and parallel algorithms for association rule mining: A comparison". Technical CS-TR-3515, University of Maryland, College Park, August 1995, pp.1-5.

[5] R. Agarwal and J.C. Shafer. Parallel mining of association rules: Design, implementation, and experience. In IBM Research Report, 1996.

[6] D. Cheung et al., " A Fast Distributed Algorithm for Mining Association Rules," Proc, 4th Int'l Conf. Parallel and Distributed Information Systems, IEEE computer Soc. Press, Los Alamitors, Calif., 1996, pp. 31-42.

[7] D. Gillick, A. Faria, and J. Denero, "Mapreduce: Distributed computing for machine learning, " 2008.[online]. Available:

[8] Jiawei Han, MichelineKamber, Simon Fraser University, A book on "Data Mining: concepts and Techniques", Academic press, Morgan Kaufmann Publishers, 2001, pp.227-240.