International Journal of Scientific Research in Computer Science, Engineering and Information Technology
© 2018 IJSRCSEIT | Volume 3 | Issue 1 | ISSN : 2456-3307

Mining Association Rules in Cloud Computing Environments using Modified Apriori Algorithm

Avinash Sharma[1], Dr. N. K. Tiwari[2]
[1]Research Scholar, Bansal Group of Institutions, Bhopal, Madhya Pradesh, India
[2]Director, Bansal Group of Institutions, Bhopal, Madhya Pradesh, India

ABSTRACT

An association rule mining helps in finding relation between the items or item sets in the given data. The performance of the algorithm was evaluated by testing it in the cloud (EC2) by increasing the number of nodes in the testing set up. The association rules are developed on the basis of the frequent item set generated from the data. The frequent item set were generated following the Apriori algorithm. As the input data and number of distinct items in the data set is large, lots of space and memory is required. Association rules are dependency rules which predict occurrence of an item based on occurrences of other items. Apriori is the best-known algorithm to mine association rules. The Apriori algorithm had a major problem of multiple scans through the entire data. It required a lot of space and time. The modification in our paper suggests that we do not scan the whole database to count the support for every attribute. This is possible by keeping the count of minimum support and then comparing it with the support of every attribute. The support of an attribute is counted only till the time it reaches the minimum support value. In this paper we use Modified Apriori algorithm to mine the data from the cloud using sector/sphere framework with association rules.

Key words: Data Mining, Cloud Computing Association Rules

I. INTRODUCTION

Data mining is the process of analyzing data from different perspectives and summarizing it into useful information - information that can be used to increase revenue, cuts costs, or both. It allows users to analyze data from many different dimensions or angles, categorize it, and summarize the relationships identified. Mining Association rule is a way to find interesting associations among large sets of data items. Using this we have determined the frequent item sets based on a predefined support [6].

By cloud we can say that it is an infrastructure that consists of services delivered through shared datacenters and appearing as a single point of access for consumers' computing needs and also provides demanded resources and/or service over the internet. Sector storage cloud is a distributed storage system that can be deployed over a wide area network and allows users to consume and download large dataset from any location with a high-speed network connection to the system. Sector automatically replicates files for the better reliability, access and availability. Sphere compute cloud is a computation service which is built on the top of the sector storage cloud. It allows developers to write certain distributed data intensive parallel applications with several simple APIs. Data locality is the key factor for the performance in the Sphere. Thus to summarize we can say that sector manages data in form of distributed indexed files, sphere processes that data using sphere processing engine that is applied parallel on every data segment managed by sector

Frequent Pattern Mining is most powerful problem in association mining. Most of the algorithms are

based on algorithm is a classical algorithm of association rule mining [2,3, 4]. Lots of algorithms for mining association rules and their mutations are proposed on basis of Apriori Algorithm [2, 3]. Most of the previous studies adopt Apriori-like algorithms, which generate-and-test candidates and improving algorithm strategy and structure. Several modifications on apriori algorithm are focused on algorithm Strategy but no one algorithm emphasis on representation of database. A simple approach is if we implement in Transposed database then result is very fast. Recently, different works proposed a new way to mine patterns in transposed databases where a database with thousands of attributes but only tens of objects [2]. In many example attribute are very large than objects or transaction In this case, mining the transposed database runs through a smaller search space. In apriori algorithm each phase is count the support of prune pattern candidate from database. No one algorithm filters or reduces the database in each pass of apriori algorithm to count the support of prune pattern candidate from database. The Apriori algorithm had a major problem of multiple scans through the entire data. It required a lot of space and time. Apriori Algorithm Scans the database too many times, When the database storing a large number of data services, the limited memory capacity, the system I/O load, considerable time scanning the database will be a very long time, so efficiency is very low. In order to overcome the drawback inherited in Apriori.

The modification in our paper suggests that we do not scan the whole database to count the support for every attribute. This is possible by keeping the count of minimum support and then comparing it with the support of every attribute. The support of an attribute is counted only till the time it reaches the minimum support value. Beyond that the support for an attribute need not be known. This provision is possible by using a variable named flag in the algorithm. As soon as flag changes its value, the loop is broken and the value for support is noted. In this paper we have discussed an Modified algorithm to mine the data from the cloud using sector/sphere framework with association rules

II. LITERATURE SURVEY

In this section, we briefly review the most related studies including frequent pattern mining algorithms and parallel and distributed algorithms for frequent pattern mining.

In 2010, Kawuu W.Lin et al. [1] proposed a set of strategies for many-task frequent pattern mining. Through empirical evaluations on various simulation conditions, the proposed strategies deliver excellent performance in terms of execution time.

In 2010, Yang Lai et al. [2] proposed a data mining framework on Hadoop using the Java Persistence API (JPA) and MySQL Cluster. The framework is elaborated in the implementation of a decision tree algorithm on Hadoop. We compare the data indexing algorithm with Hadoop MapFile indexing, which performs a binary search, in a modest cloud environment. The results show the algorithm is more efficient than naïve MapFile indexing. They compare the JDBC and JPA implementations of the data mining framework. The performance shows the framework is efficient for data mining on Hadoop.

In 2010, Jiabin Deng et al. [3] propose about the use of Power-law Distributions and Improved Cubic Spline Interpolation for multi-perspective analysis of shareware download frequency. The tasks include data mining the usage patterns and to build a mathematical model. Through analysis and checks, in accordance with changes to usage requirements, our proposed methods will intelligently adjust the data redundancy of cloud storage. Thus, storage resources are fine tuned and storage efficiency is greatly enhanced

In 2011, Lingjuan Li et al. [4] proposed a strategy of mining association rules in cloud computing

environment is focused on. Firstly, cloud computing, Hadoop, MapReduce programming model, Apriori algorithm and parallel association rule mining algorithm are introduced. Then, a parallel association rule mining strategy adapting to the cloud computing environment is designed. It includes data set division method, data set allocation method, improved Apriori algorithm, and the implementation procedure of the improved Apriori algorithm on MapReduce. Finally, the Hadoop platform is built and the experiment for testing performance of the strategy as well as the improved algorithm has been done.

In 2011, T.R. Gopalakrishnan Nair et al. [5] presents a specific method of implementing kmeans approach for data mining in such scenarios. In this approach data is geographically distributed in multiple regions formed under several virtual machines. The results show that hierarchical virtual k-means approach is an efficient mining scheme for cloud databases.

In 2011, Lingjuan Li et al. [6] Focus on the strategy of mining association rules in cloud computing environment. Firstly, cloud computing, Hadoop, Map Reduce programming model, Apriori algorithm and parallel association rule mining algorithm are introduced. Then, a parallel association rule mining strategy adapting to the cloud computing environment is designed. It includes data set division method, data set allocation method, improved Apriori algorithm, and the implementation procedure of the improved Apriori algorithm on Map Reduce. Finally, the Hadoop platform is built and the experiment for testing performance of the strategy as well as the improved algorithm has been done.

In 2011, Fabrizio Marozzo et al. [7] present a Data Mining Cloud App framework that supports the execution of parameter sweeping data mining applications on a Cloud. The framework has been implemented using the Windows Azure platform, and evaluated through a set of parameter sweeping clustering and classification applications. The experimental results demonstrate the effectiveness of the proposed framework, as well as the scalability that can be achieved through the parallel execution of parameter sweeping applications on a pool of virtual servers.

III. DATA MINING IN CLOUD COMPUTING

Data mining techniques and applications are very much needed in the cloud computing paradigm. As cloud computing is penetrating more and more in all ranges of business and scientific computing, it becomes a great area to be focused by data mining.

"Cloud computing denotes the new trend in Internet services that rely on clouds of servers to handle tasks. Data mining in cloud computing is the process of extracting structured information from unstructured or semi-structured web data sources.

The data mining in Cloud Computing allows organizations to centralize the management of software and data storage, with assurance of efficient, reliable and secure services for their users." [7]

As Cloud computing refers to software and hardware delivered as services over the Internet, in Cloud computing data mining software is also provided in this way.

The main effects of data mining tools being delivered by the Cloud are:

- ✓ the customer only pays for the data mining tools that he needs – that reduces his costs since he doesn't have to pay for complex data mining suites that he is not using exhaustive;
- ✓ the customer doesn't have to maintain a hardware infrastructure, as he can apply data mining through a browser – this means that he has to pay only the costs that are generated by using Cloud computing.

Using data mining through Cloud computing reduces the barriers that keep small companies from benefiting of the data mining instruments.

"Cloud Computing denotes the new trend in Internet services that rely on clouds of servers to handle tasks. Data mining in cloud computing is the process of extracting structured information from unstructured or semi-structured web data sources.

The data mining in Cloud Computing allows organizations to centralize the management of software and data storage, with assurance of efficient, reliable and secure services for their users." [6]

The implementation of data mining techniques through Cloud computing will allow the users to retrieve meaningful information from virtually integrated data warehouse that reduces the costs of infrastructure and storage

IV. PROBLEM IDENTIFICATION

Association rule mining is a popular and well researched area for discovering interesting relations between variables in large databases for Cloud Computing Environment. We have to analyze the coloring process of dyeing unit using association rule mining algorithms using frequent patterns. These frequent patterns have a confidence for different treatments of the dyeing process. These confidences help the dyeing unit expert called dyer to predict better combination or association of treatments.

Various algorithms are used for the coloring process of dyeing unit using association rules. For example. LRM,FP Growth Method., H-Mine and Aprori algorithm But these algorithm significantly reduces the size of candidate sets . However, it can suffer from three-nontrivial costs:

1. Generating a huge number of candidate sets, and
2. Repeatedly scanning the database and checking the candidates by pattern matching.
3. It take more time for generate frequent item set.
4. The large databases can not be executed efficiently in H-Mine and LRM algorithms,

We have to proposed such that algorithm that it has a very limited and precisely predictable main memory cost and runs very quickly in memory-based settings. it can be scaled up to very large databases using database partitioning and to identify the better dyeing process of dyeing unit.

V. PROPOSED ALGORITHM

The Apriori algorithm had a major problem of multiple scans through the entire data. It required a lot of space and time. The modification in our paper suggests that we do not scan the whole database to count the support for every attribute. This is possible by keeping the count of minimum support and then comparing it with the support of every attribute. The support of an attribute is counted only till the time it reaches the minimum support value. Beyond that the support for an attribute need not be known. This provision is possible by using a variable named flag in the algorithm. As soon as flag changes its value, the loop is broken and the value for support is noted. The pseudo code for the proposed algorithm is as follows:

Input : Database, D, of transactions;
Minimum support threshold, min_sup

Output : L, frequent itemsets in D

Method :

```
1) L(1)= find_frequent _1-itemsets(D);
2) For each transaction t belongs to D
3) count_items= count_items(t);
4) For (k=2; L(k-1)!=null; k++)
5) {
6) C(k)= apriori_gen(L(k-1, min_sup);
7) flag=1;
8) For each transaction t belonging to D
   Where count_items>=k
9) {
10) If (flag==1)
11) {
12) c=subset(C(k),t);
```

13) c.count++;
14) if (c.count==min_sup)
15) flag=0;
16) }
17) if (flag==0)
18) Exit from loop
19) }
20) L(k)={c.count=min_sup}
21) }
22) return L=U(k) L(k);

VI. EXPERIMENTAL RESULT AND ANALYSIS

Now, we compare the association rules mining algorithms on the whole data set with 5000 data set The Computation time results for the clustering algorithms are shown in Table 1 respectively.

Now we implement the association rules in web usage mining that Apriori algorithm is more efficient which takes less time, less memory and hence results in high efficiency. The experimental results shows improvement in generation of candidate sets, results in reduced number of data base scan, and also the time and space consumption. we calculate support and confidence

$$support = \frac{(X \cup Y).count}{n}$$

$$confidence = \frac{(X \cup Y).count}{X.count}$$

Table 1. Support and Confidence value of modified Apriori algorithm

Database Size	Modified apriori algorithm	
	Support	Confidence
200	52	56
400	65	69
600	66	70
800	71	79
1000	69	78

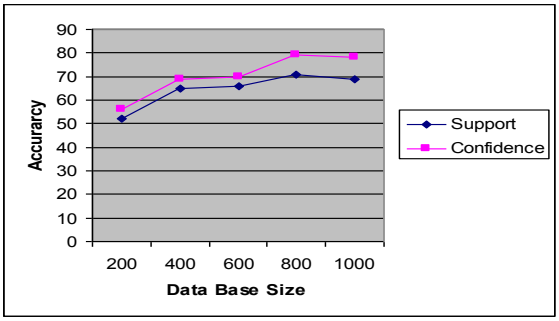

Figure 1. Graph for Support and Confidence of modified Apriori Algorithm

VII. CONCLUSION

In this paper we have attempted to give a new perspective algorithm with the eye of a modified apriori algorithm.

This algorithm is better than both of the previous methods, i.e., FPGrowth tree algorithm and TPFP algorithm. This method works perfectly for data that has been supervised, i.e., data whose classes are already known. But if the classes are not known already, then we can first take any attributes as prominent attributes and test them for modified apriori. Also, the data taken in this example is discrete and this algorithm works on numeric data.

VIII. REFERENCES

[1]. KawuuW.Lin,Yu-ChinLuo ," Efficient Strategies for Many-task Frequent Pattern Mining in Cloud Computing Environments",2010 IEEE.

[2]. Yang Lai , Shi ZhongZhi ," An Efficient Data Mining Framework on Hadoop using Java Persistence API" , 2010 10th IEEE International Conference on Computer and Information Technology (CIT 2010).

[3]. Jiabin Deng, JuanLi Hu, Anthony Chak Ming LIU, Juebo Wu, "Research and Application of Cloud Storage",2010 IEEE.

[4]. Lingjuan Li , Min Zhang , "The Strategy of Mining Association Rule Based on Cloud Computing", 2011 IEEE.

[5]. T.R. Gopalakrishnan Nair, K.Lakshmi Madhuri , "Data Mining Using Hierarchical Virtual KMeans Approach Integrating Data Fragments In Cloud Computing Environment",2011 IEEE.

[6]. L. J. Li and M. Zhang, "The strategy of mining association rule based on cloud computing," in Proc. 2011 International Conference on Business Computing and Global Informatization.

[7]. F. Marozzo, D. Talia, and P. Trunfio, "A cloud framework for parameter sweeping data mining applications," in Proc. 2011 Third IEEE International Conference on Coud Computing Technology and Science.

[8]. R. Agrawal, R. Srikant, Mining Sequential Patterns, in: Proc. of the 11th Int'l Conf. on Data Engineering, 1995, pp. 3-14.

[9]. R. J. Bayardo, Jr., Brute-force mining of high-confidence classification rules. In Proceedings of the 3rd international conference on knowledge discovery and data mining (KDD'97), Newport Beach, California, USA.

[10]. M. Ester, H.-P. Kriegel, J. Sander, and X. Xu, 1996, "A Density-Based Algorithm for Discovering Clusters in Large Spatial Databases with Noise", In Proc. 2nd Int. Conf. on Knowledge Discovery and Data Mining, Portland, OR, AAAI Press, pp. 226-231.

[11]. G. Grahne and J. Zhu, 2003, "Efficiently Using Prefix-trees in Mining Frequent Itemsets", In Proc. of the IEEE ICDM Workshop on Frequent Itemset Mining Implementations.

[12]. J. Han, J. Pei, and Y. Yin, 2000, "Mining Frequent Patterns without Candidate Generation", In Proc. of the ACM SIGMOD Int. Conf. on Management of Data, pp.1-12.

[13]. A. Javed, and A. Khokhar, 2004, "Frequent Pattern Mining on Message Passing Multiprocessor Systems", Distributed and Parallel Databases, vol. 16, pp. 321–334.

[14]. K. W. Lin, Y.-C. Luo, 2009, "A Fast Parallel Algorithm for Discovering Frequent Patterns", GRC '09. IEEE Int. Conf. on Granular Computing, pp. 398 – 403.

[15]. J. Zhou and K.-M. Yu, 2008, "Tidset-based Parallel FP-tree Algorithm for the Frequent Pattern Mining Problem on PC Clusters", Lecture Notes in Computer Science 5036, pp. 18- 28.

[16]. J. Zhou and K.-M. Yu, 2008, "Balanced Tidset-based Parallel FP-tree Algorithm for the Frequent Pattern Mining on Grid System", Fourth Int. Conf. on Semantics, Knowledge and Grid, pp. 103-108.

[17]. R. Agrawal and R. Srikant. Quest Synthetic Data Generator. IBM Almaden Research Center, San Jose, California, http://www.almaden.ibm.com/cs/quest/syndata.html.

[18]. R. Agrawal, T. Imielinski, and A. Swami, 1993, "Mining association rules between sets of items in large databases", In Proc. of the 1993 ACM-SIGMOD Int. Conf. on management of data (SIGMOD'93), pp. 207–216.