Modified Apriori Algorithm For Predefind Support And Confidence In Cloud Computing Environment For Frequent Pattern Mining

Ms. Roshani Parate
M.Tech Scholor NRI College Bhopal

Prof. Sitendra Tamarkar
M.Tech. Coordinator

Abstract

Cloud computing has demonstrated that processing very large datasets over commodity clusters can be done by giving the right programming model. Cloud can be meant as an infrastructure that provides resources and/or service over the internet. A cloud can be a storage cloud that provides block or file based storage service or it can be a compute cloud that provides computational services. Data mining is the process of analyzing data from different perspectives and summarizing it into useful information. Mining association rules is one of the most important aspects in data mining. 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. using sector/sphere framework with association rules.

.

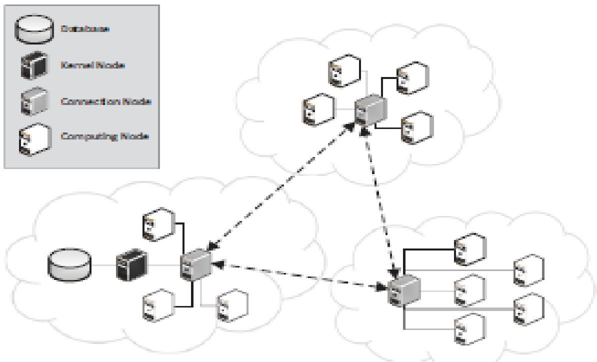
1. Introduction

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 Application program interfaces. 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 algorithm FD-Mine[9]. The main characteristics of it contain compressing the whole FP-tree for preserving data privacy and abating the network latency, and addressing a better dispatching workset strategy for load balancing than BTPtree[11]which extends TPFP-tree.

In system architecture of FD-Mine, there is a kernel



3. 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

4. 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.

5. 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. Up to 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 :

```
L(1)= find_frequent _1-itemsets(D);
For each transaction t belongs to D
count_items= count_items(t);
For (k=2; L(k-1)!=null; k++)
{
C(k)= apriori_gen(L(k-1, min_sup);
flag=1;
For each transaction t belonging to D
   Where count_items>=k
{
If (flag==1)
{
c=subset(C(k),t);
c.count++;
if (c.count==min_sup)
flag=0;
}
if (flag==0)
Exit from loop
}
L(k)={c.count=min_sup}
}
return L=U(k) L(k);
```

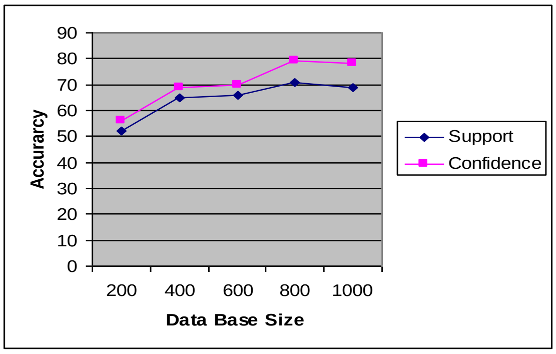
6. Experimental ouput

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

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

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

Table1: Support and Confidence value of modified Apriori algorithm



7. 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., FP Growth 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.

References

1. R. Agrawal, R. Srikant, Mining Sequential Patterns, in: Proc. of the 11th Int'l Conf. on Data Engineering, 1995, pp. 3-14.
2. 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.
3. 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.
4. 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.
5. 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.
6. A. Javed, and A. Khokhar, 2004, "Frequent Pattern Mining on Message Passing Multiprocessor Systems", Distributed and Parallel Databases, vol. 16, pp. 321–334.
7. 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.
8. 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.
9. 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.
10. 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.
11. 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.