An Approach to Mine Frequent Itemsets in Cloud Using Apriori and FP-tree Approach

[1] Harneet Khurana, [2] Kailash Bahl

[1] Research Scholar, M.Tech (CSE), Patiala Institute of Engineering and Technology for Women
Patiala, Punjab, India

[2] Dean Academics, Patiala Institute of Engineering and Technology for Women
Patiala, Punjab, India

Abstract - Cloud computing has become a big name in present era. It has proved to be a great solution for storing and processing huge amount of data. It provides us demand, scalable, pay-as-you go compute and storage capacity. Data mining techniques implemented with cloud computing paradigm are very useful to analyze big data on clouds. In our dissertation we have used association rule mining as a data mining technique. In particular we have used Apriori algorithm for association rule mining. It has been observed that the original Apriori algorithm was designed for sequential computation so directly using it for parallel computation doesn't seems a good idea. So we have improved the Apriori algorithm (FP Growth) so as to suit it for parallel computation platform. We have used CloudSim Simulator for cloud computing.

***Keywords* - Data Mining, Cloud Computing, Association Rule Mining in Clouds, Apriori Algorithm, FP-Growth Algorithm.**

1. Introduction

With the increase in Information Technology, the size of the databases created by the organizations due to the availability of low-cost storage and the evolution in the data capturing technologies is also increasing. These organization sectors include retail, petroleum, telecommunications, utilities, manufacturing, transportation, credit cards, insurance, banking and many others, extracting the valuable data, it necessary to explore the databases completely and efficiently. Knowledge discovery in databases (KDD) helps to identifying precious information in such huge databases. This valuable information can help the decision maker to make accurate future decisions. KDD applications deliver measurable benefits, including reduced cost of doing business, enhanced profitability, and improved quality of service. Therefore Knowledge Discovery in Databases has become one of the most active and exciting research areas in the database community.

Cloud computing can be defined as the use of computing resources that are delivered as a service over a network. With traditional computing paradigms we run the software and store data on our computer system. These files could be shared in a network. The importance of cloud computing lies in the fact that the software are not run from our computer but rather stored on the server and accessed through internet. Even if a computer crashes, the software is still available for others to use. The concept of cloud computing has developed from clouds. A cloud can be considered as a large group of interconnected computers which can be personal computers or network servers; they can be public or private. The concept of cloud computing has spread rapidly through the information technology industry. The ability of organizations to tap into computer applications and other software via the cloud and thus free themselves from building and managing their own technology infrastructure seems potentially irresistible. In fact some companies providing cloud services have been growing at double digit rates despite the recent economic downturn.

Cloud Mining can be considered as a new approach to apply Data Mining. There is a lot of data and unfortunately this huge amount of data is difficult to mine and analyze in terms of computational resources. With the cloud computing paradigm the data mining and analysis can be more accessible and easy due to cost effective computational resources. Here we have discussed the usage of cloud computing platforms as a possible solution for mining and analyzing large amounts of data.

2. Overview

The overview of important terms is as follows:

2.1 Data Mining

Data mining, or knowledge discovery, is the computer-assisted process of digging through and analyzing enormous sets of data and then extracting the meaning of the data. Data mining tools predict behaviors and future trends, allowing businesses to make proactive, knowledge-driven decisions. Data mining tools can answer business questions that traditionally were too time consuming to resolve. They scour databases for hidden patterns, finding predictive information that experts may miss because it lies outside their expectations. While large-scale information technology has been evolving separate transaction and analytical systems, data mining provides the link between the two. Data mining software analyzes relationships and patterns in stored transaction data based on open-ended user queries. Several types of analytical software are available: statistical, machine learning, and neural networks. Generally, any of four types of relationships are sought:

- **Classes**: Stored data is used to locate data in predetermined groups. For example, a restaurant chain could mine customer purchase data to determine when customers visit and what they typically order. This information could be used to increase traffic by having daily specials.
- **Clusters**: Data items are grouped according to logical relationships or consumer preferences. For example, data can be mined to identify market segments or consumer affinities.
- **Associations**: Data can be mined to identify associations. The beer-diaper example is an example of associative mining.
- **Sequential patterns**: Data is mined to anticipate behavior patterns and trends. For example, an outdoor equipment retailer could predict the likelihood of a backpack being purchased based on a consumer's purchase of sleeping bags and hiking shoes.

2.2 Cloud Computing

Cloud computing can be defined as the use of IT resources (such as software’s, platforms, storage etc) that are delivered as a service over a network. With traditional computing paradigms we run the software and store data on our computer system. These files could be shared in a network. The importance of cloud computing lies in the fact that the software are not run from our computer but rather stored on the server and accessed through internet. Even if a computer crashes, the software is still available for others to use. The concept of cloud computing has developed from clouds. A cloud can be considered as a large group of interconnected computers which can be personal computers or network servers; they can be public or private. The concept of cloud computing has spread rapidly through the information technology industry. The ability of organizations to tap into computer applications and other software via the cloud and thus free themselves from building and managing their own technology infrastructure seems potentially irresistible. In fact some companies providing cloud services have been growing at double digit rates despite the recent economic downturn.

2.3 Apriori Algorithm

Apriori is the very first algorithm for mining frequent patterns. It was given by R Agrawal and R Srikant in 1994.It works on horizontal layout based database. It is based on Boolean association rules which uses generate and test approach. It uses BFS (breadth first search). Apriori uses frequent k itemsets to find a bigger itemset of k+1 items. In Apriori support count for each item is given, the algorithm first scan the database to find out all frequent items based on support. The calculation of frequency of an item is done by counting it’s occurrence in all transactions [6]. All infrequent items are dropped.
Apriori property: All subsets of a frequent itemsets which are non empty are also frequent.

Apriori follows two steps approach:

In the first step it joins two itemsets which contain k-1 common items in kth pass. The first pass starts from the single item, the resulting set is called the candidate set Ck. In the second step the algorithm counts the occurrence of each candidate set and prunes all infrequent itemsets. The algorithm ends when no further extension found.

2.4 FP Growth Algorithm

Frequent pattern growth also labeled as FP growth is a tree based algorithm to mine frequent patterns in database the idea was given by (han et. al. 2000) .It is applicable to projected type database. It uses divide and conquer method. In it no candidate frequent itemset is needed rather frequent patterns are mined from fp tree. In the first step a list of frequent itemset is generated and sorted in their decreasing support order. This list is represented by a structure called node. Each node in the fp tree, other than the root node, will contain the item name, support count, and a pointer to link to a node in the tree that has the same item name. These nodes are used to create the fp tree. Common prefixes can be shared during fp tree construction. The paths from root to leaf nodes are arranged in non increasing order of their support.

Once the fp tree is constructed then frequent patterns are extracted from the fp tree starting from the leaf nodes. Each prefix path subtree is processed recursively to mine frequent itemsets. FP Growth takes least memory because of projected layout and is storage efficient. A variant of fp tree is conditional FP tree that would be built if we consider transactions containing a particular itemset and then removing that itemset from all transactions. Another variant is parallel fp growth (PFP) that is proposed to parallelize the fp tree on distributed machines. FP Growth is improved using prefix-tree-structure, Grahne and Zhu.

3. Proposed System

We will create a virtual cloud environment in Java using CloudSim simulator. Then in second step we will collect and preprocess the data. Then we will apply and analyze the performance of apriori and improved apriori algorithm (FPGROWTH) on the dataset.

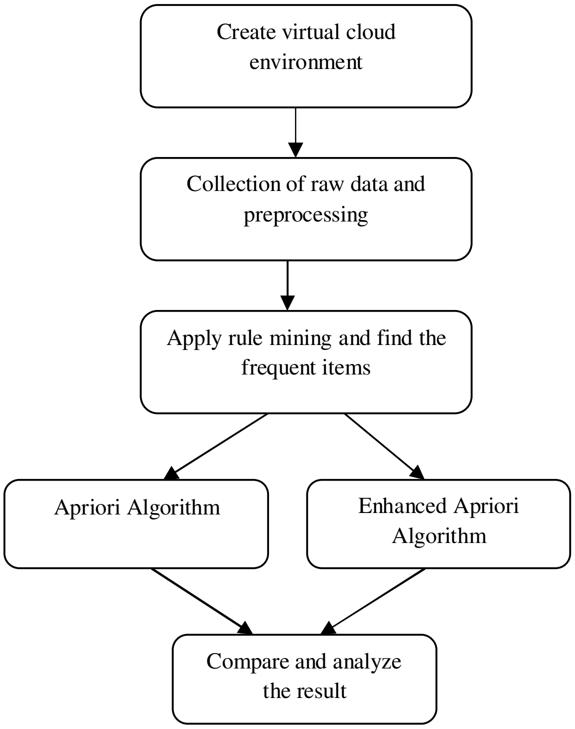
Create virtual cloud environment
Collection of raw data and preprocessing
Apply rule mining and find the frequent items
Apriori Algorithm
Enhanced Apriori Algorithm
Compare and analyze the result

Fig. 1 Proposed System for cloud mining

4. Methodology

Various data mining techniques have been implemented in cloud computing. The Apriori algorithm is a famous algorithm for association rule mining. But the existing implementation used the original Apriori for cloud computing paradigm. Some have tried parallelism but have failed to reduce number of steps in Apriori algorithm. Using original Apriori for cloud paradigm doesn't make a good choice because the original Apriori algorithm was designed for the sequential computing. So in this project we will use improved apriori algorithm (FP Growth) on the cloud platform.

5. Conclusion

Cloud computing is an architecture which is known for its powerful capability of computation and storage and resource sharing. These features make cloud computing favorable to data mining service in network environment. We have discussed association rule mining in cloud environment and various parallel and distributed mining algorithms. Data mining on cloud computing paradigm can benefit us to a great extent. That is why we have implemented data mining technique on cloud platform. Out of many data mining techniques we have studied association rule mining technique in this paper. More specifically we have association rule mining in cloud computing environment.

Acknowledgments

Thanks to my Guide and family members who always support, help and guide me during my dissertation.

References

[1] Jiawei Han Micheline Kamber, Data Mining concepts and techniques, 2nd Ed.

[2] Sanjeev Rao, Priyanka Gupta, ―Implementing Improved Algorithm Over APRIORI Data Mining Association Rule Algorithm‖, ISSN: 0976-8491 (Online) | ISSN: 2229-4333 (Print) IJCST Vol. 3, Issue 1, Jan. - March 2012.

[3] Bhagyashree Ambulkar, Vaishali Borkar, ―Data Mining in Cloud Computing‖, Proceedings published by International Journal of Computer Applications® (IJCA)ISSN: 0975 – 8887.

[4] Deepak Garg et. al. "Comparative Analysis of Various Approaches Used in Frequent Pattern Mining" (IJACSA) International Journal of Advanced Computer Science and Applications, Special Issue on Artificial Intelligence.

[5] Han J. and M. Kamber, "Data Mining Concepts and Techniques", Morgan Kaufmann publishers, 2nd Edition.

[6] Jing Ding, Shanlin Yang, ―Classification Rules Mining Model with Genetic Algorithm in Cloud Computing‖, International Journal of Computer Applications (0975 – 888), Volume 48– No.18, June 2012.

[7] Goswami D.N et. al. "An Algorithm for Frequent Pattern Mining Based On Apriori " (IJCSE) International Journal on Computer Science and Engineering Vol. 02, No. 04, 2010, Pp. 942-947.

[8] Yunhong Gu and Robert L. Grossman "Sector and Sphere: The Design and Implementation of a High Performance Data Cloud".

[9] Ramakrishnan Srikant and Rakesh Agrawal "Mining Generalized Association Rules".

[10] M. Bramer. Principles of Data Mining. Springer, 2007.

Biography

First Author Harneet Khurana, Research Scholar, is pursuing M.Tech in Computer Science and Engineering in Patiala Institute of Engineering And Technology for Women, Patiala, Punjab, India.

Second Author Kailash Bahl, Dean Academics in Patiala Institute Of Engineering and Technology for Women, Patiala, Punjab, India.