A Review on Data Mining Algorithms for Cloud Services

Pradhumn Soni[1] Mani Butwall[2]
[1,2]Assistant Professor,
[1,2]Department of Computer Science & Engineering
[1,2]MIT, Mandsaur, India

Abstract— Data and information are basis for today's industrial and day to day applications of almost every domain. The volume of data for most of human friendly applications is generally enormous in nature and thus it is mandatory to scale them in pre-defined clusters based on their properties. Data mining has always interested the researchers to characterize the data and is integrated in applications like: Statistics, Machine Learning, Artificial Intelligence, pattern recognition etc. Data mining applications can derive much demographic information concerning customers that was previously not known or hidden in the data. We have recently seen an increase in data mining techniques targeted to such applications as fraud detection, identifying criminal suspects, and prediction of potential terrorists. By and large, data mining systems that have been developed to data for clusters, distributed clusters and grids have assumed that the processors are the scarce resource, and hence shared. When processors become available, the data is moved to the processors. This paper surveys some data mining methods in their actual nature and modifications made in their algorithms for better output in their performance.

Key words: Data mining, Apriori Algorithm

I. INTRODUCTION

The amount of data kept in computer files is growing at a phenomenal rate. The data mining field offers to discover unknown information. Data mining is often defined as the process of discovering interesting patterns from the large amount of data stored in repositories. Association rule mining, clustering, classification etc are some of the important techniques of data mining. Data mining is usually used in market basket analysis and can be applied in various fields. The problem of mining association rules over market basket data [1]. Market basket data contain a set of transactions, wherever each transaction is a set of objects.

Association rule mining is used to find a set of association rules of form $X \rightarrow Y$, where X and Y are a disjoint set of items. For example, customers who buy shoes also buy socks: shoes → socks. Support and confidence are used to determine the importance of rules. Support of a rule is defined as $p(X \cup Y)$. While confidence of a rule $X \rightarrow Y$ is defined as $p\,(Y/X)$. Goal of association rule mining is to find all rules that satisfy user-given minimum support and minimum confidence threshold. Applications of association rule mining include customer behavior analysis, bioinformatics, intrusion detection, association classification etc.

Data mining techniques like clustering and association rule mining can be applied on files which contains a huge amount of information to discover interesting information. This interesting information when analyzed can reveal vital information required for data improvement and thereby attract more users to access the data. Originally, association rule mining algorithms were applied for market basket analysis which contained transaction data. The transaction data may include many records of which each record has a transaction id and a list of items purchased during that transaction.

Cloud computing can be considered a good platform for association rule mining because usually the input data is very large and distributed in nature. The beauty of cloud computing is that it can provide centralized storage and processing easily. In addition we have to pay for just what we use so it also cuts down the cost.

Hadoop is a programming framework that supports the processing of large data in a distributed computing environment. It provides inexpensive and reliable storage for analyzing structured and unstructured data. Hadoop consists of two modules: (1) Hadoop Distributed File System (HDFS) for reliable storage and (2) Map/Reduce for high-performance parallel data processing. However converting sequential algorithms to parallel form can be challenging. Thus implementation of association rule mining on Hadoop may not be a good choice.

Cloud computing combined with data mining can provide powerful capacities of storage and computing and an excellent resource management [2]. Due to the explosive data growth and amount of computation involved in data mining, an efficient and high-performance computing is very necessary for a successful data mining application. Data mining in the cloud computing environment can be considered as the future of data mining because of the advantages of cloud computing paradigm. Cloud computing provides greater capabilities in data mining and data analytics [3]. The major concern about data mining is that the space required by the operations and item sets is very large. But if we combine the data mining with cloud computing we can save a considerable amount of space [4]. This can benefit us to a great extent.

There are certain issues associated with data mining in the cloud computing. The major issue of data mining with cloud computing is security as the cloud provider has complete control on the underlying computing infrastructure [4]. Special care has to be taken so as to ensure the security of data under cloud computing environment.

II. ASSOCIATION RULE MINING

Association rule mining finds interesting associations and/or correlation relationships among large sets of data items. Association rules show attributes value conditions that occur frequently together in a given dataset. A typical and widely-used example of an association rule mining is Market Basket Analysis.

For example, data are collected using bar-code scanners in supermarkets. Such market basket databases consist of a large number of transaction records. Each record lists all items bought by a customer on a single purchase transaction. Managers would be interested to know if certain

groups of items are consistently purchased together. They could use this data for adjusting store layouts, for cross-selling, for promotions, for catalogue design and to identify customer segments based on buying patterns.

Association rules provide information of this type in the form of "if-then" statements. These rules are computed from the data and, unlike the if-then rules of logic, association rules are probabilistic in nature.

In addition to the antecedent (the "if" part) and the consequent (the "then" part), an association rule has two numbers that express the degree of uncertainty about the rule. In association analysis the antecedent and consequent are sets of items (called itemsets) that are disjoint (do not have any items in common).

The first number is called the support for the rule. The support is simply the number of transactions that include all items in the antecedent and consequent parts of the rule. The support is sometimes expressed as a percentage of the total number of records in the database.

The other number is known as the confidence of the rule. Confidence is the ratio of the number of transactions that include all items in the consequent as well as the antecedent (namely, the support) to the number of transactions that include all items in the antecedent.

A. Association Rule Problem:

A formal statement of the association rule problem is [Agrawal1993] [Cheung1996c]:

1) Definition 1:

Let $I = \{I_1, I_2, \ldots, I_m\}$ be a set of m distinct attributes, also called *literals*. Let D be a database, where each record (tuple) T has a unique identifier, and contains a set of items such that $T \subseteq I$ An association rule is an implication of the form $X \Rightarrow Y$, where $X, Y \subset I$, are sets of items called itemsets, and $X \cap Y = \phi$. Here, X is called antecedent, and Y consequent. Two important measures for association rules, support (s) and confidence (α), can be defined as follows.

2) Definition 2:

The support (s) of an association rule is the ratio (in percent) of the records that contain $X \cup Y$ to the total number of records in the database. Therefore, if we say that the support of a rule is 5% then it means that 5% of the total records contain $X \cup Y$. Support is the statistical significance of an association rule. Grocery store managers probably would not be concerned about how peanut butter and bread are related if less than 5% of store transactions have this combination of purchases. While a high support is often desirable for association rules, this is not always the case. For example, if we were using association rules to predict the failure of telecommunications switching nodes based on what set of events occur prior to failure, even if these events do not occur very frequently association rules showing this relationship would still be important.

3) Definition 3:

For a given number of records, confidence (α) is the ratio (in percent) of the number of records that contain $X \cup Y$ to the number of records that contain X.

Thus, if we say that a rule has a confidence of 85%, it means that 85% of the records containing X also contain Y. The confidence of a rule indicates the degree of correlation in the dataset between X and Y. Confidence is a measure of a rule's strength. Often a large confidence is required for association rules. If a set of events occur a small percentage of the time before a switch failure or if a product is purchased only very rarely with peanut butter, these relationships may not be of much use for management. Mining of association rules from a database consists of finding all rules that meet the user-specified threshold support and confidence.

An another approach was used in [5], in which candidate set is not created for getting the frequent item set, rather it creates a comparatively compact tree-structure that improves the multi-scan problem and improves the candidate itemset generation.

III. APRIORI ALGORITHM

Apriori is a classic algorithm which is proposed by Agrawal & Srikant for frequent itemset mining and association rule learning over transactional databases. It continues by recognizing the frequent individual items in the database and extending them to larger and larger itemsets as long as those itemsets appear sufficiently often in the database. The frequent itemsets determined by Apriori can be used to determine association rules which highlight general trends in the database. This has applications in domains such as market basket analysis. The problem is that we are given a set of items and a large collection of transactions which are sets (baskets) of items. The task is to discover interactions between the containments of various items within those baskets.

It is an iterative approach and there are two steps in the each iteration. The first step produces a set of candidate itemsets. Then, in the second step we count the occurrence of each candidate set in the database and prune all disqualified candidates (i.e. all infrequent itemsets). Apriori uses two pruning technique, first on the bases of support count (should be greater than user specified support threshold) and second for an item set to be frequent, all its subsets should be in last frequent itemset. The iterations begin with size 2 itemsets and the size is incremented after the each iteration. The algorithm is founded on the closure property of frequent itemsets: if a set of items is repeated, then all its proper subsets are also repeated.

The Apriori Algorithm as described in the [6]. The pseudo code for the algorithm is given below for a transaction database T, and a support threshold of ϵ.

Initialize: k := 1, C_1 = all the 1- itemsets;
read the database to count the support of C_1 to determine L_1.
L_1 := {frequent 1- itemsets};
k:=2; //k represents the pass number//
while ($L_{k-1} \neq \emptyset$) do
begin
C_k := gen_candidate_itemsets with the given L_{k-1}
prune(C_k)
for all transactions $t \in T$ do
increment the count of all candidates in C_K that are contained in t;
L_k := All candidates in C_k with minimum support ;
k := k + 1;
end
Answer := $\cup_k L_k$

IV. HADOOP

Hadoop is an open source software framework that helps in the distributed processing of large data sets across clusters. Hadoop is based on a software framework where an application is divided into smaller parts and these parts can be run on any node in the cluster (figure 1). Hadoop is highly scalable and a fault tolerant framework. It can scale from a single to thousands of machines. Hadoop has two main modules namely, Mapreduce and HDFS.

MapReduce is the software framework that is used to write applications that process large amount of data across the cluster. MapReduce works by splitting the input data into smaller independent parts which are processed by the mappers in parallel fashion. The outputs of the mappers are then fed to the reducers. Through Hadoop the tasks can be scheduled, monitored and can be re-executed when failed.

The other module of Hadoop is HDFS (Hadoop Distributed File system). It is a file system that is used for data storage and it spans all the nodes in a cluster. It is used to link or connect together the file systems on local nodes and combines them into one big file system. HDFS also replicates data on multiple nodes to assure reliability in case of failures.

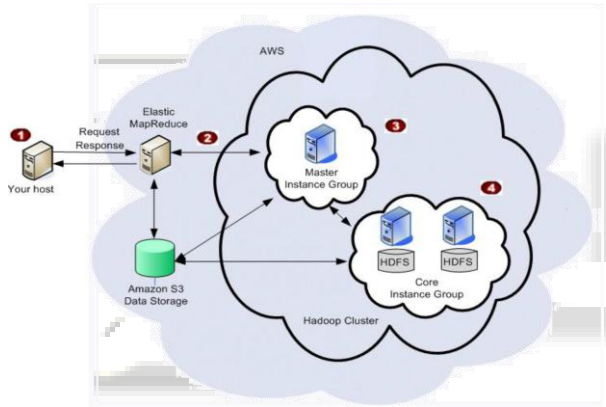

Fig. 1: Job flow in the cloud for Hadoop [7]

REFERENCES

[1] R. Agrawal, T. Imielinski, and A. N. Swami, "Mining association rules between sets of items in large databases", SIGMOD, pp. 207–216, 1993
[2] Ling Juan Li, min Zhang, ―The strategy of mining association rule based on cloud computing‖, International conference on business computing and global information, 2011
[3] Bhagyashree Ambulkar, Vaishali Borkar, ―Data Mining in Cloud Computing‖, Proceedings published by International Journal of Computer Applications® (IJCA)ISSN: 0975 – 8887.
[4] Astha Pareek, Dr. Manish Gupta, ―Review of Data Mining Techniques in Cloud Computing Database‖, International Journal of Advanced Computer Research (ISSN (print): 2249-7277 ISSN (online): 2277-7970), Volume 2 Number 2 June 2012
[5] J. Han, J. Pei, and Y. Yin, "Mining frequent patterns without candidate generation", In ACM-SIGMOD, Dallas, 2000
[6] R. Agrawal and R. Srikant, "Fast algorithms for mining association rules", In Proc. 1994 Int. Conf. Very Large Data Bases, pages 487-499, Santiago, Chile, September 1994
[7] http://docs.amazonwebservices.com

