An Efficient Cloud Data Mining (CDM) Algorithm for Frequent Pattern Mining in Cloud Computing Environment

Dheresh Soni, Atish Mishra, and Hitesh Gupta

Abstract—Cloud computing has been acknowledged as one of the prevailing models for providing IT capacities. The computing paradigm that comes with cloud computing has incurred great concerns on the security of data, especially the integrity and confidentiality of data, as cloud service providers may have complete control on the computing infrastructure that underpins the services. In this paper we want to generalize the formulation of data mining techniques with cloud computing environment. In data mining we want to find useful patterns with different methodology. The main issue with data mining techniques is that the space required for the item set and there operations are very huge. If we combine data mining techniques with cloud computing environment, then we can rent the space from the cloud providers on demand. This solution can solve the problem of huge space and we can apply data mining techniques without taking any consideration of space. This paper basically survey and analyze the utility for solving the above situation.

Index Terms—Cloud computing, data mining, frequent pattern, cloud storage.

I. INTRODUCTION

The term “Cloud computing” describes it as a system platform or a kind of software application. First, a system platform means, based on real time, it can dynamically proviso, configure, re-configure and de-proviso a system. In a cloud computing platform, server is a physical server or a virtual server. High end cloud computing generally includes other computation resources.

Cloud Computing [1], [2] is a new business model. It distributes the computing tasks to the resource pool constituted of a large number of computers, so that a variety of application systems can obtain computing power, storage space and a variety of software services on demand. The novelty of the Cloud Computing is that it almost provides unlimited cheap storage and computing power. This provides a platform for the storage and mining of mass data.

The role of data analytics increases in several request domains to cope with the big amount of captured data. Cloud computing has become one of the key considerations both in academic world and industry. Cheap, apparently infinite computing resources that can be allocated approximately right away and pay-as-you-go pricing schemes are some of the reasons for the success of cloud computing. We discuss few aspects of cloud computing and also there area. We also propose a novel approach which is cloud computing mapping and management through class and object hierarchy. In this approach we first design a cloud environment where we can analyze several object oriented aspects based on some assumptions. Then we deduce message passing behavior through a backup files based on the properties of object orient like class and object.

Association rule mining is an important research topic of data mining; its task is to find all subsets of items which frequently occur, and the relationship between them. Association rule mining has two main steps: the establishment of frequent item sets and the establishment of association rules. Apriori algorithm [3] is the most classic and most widely used algorithm for mining frequent item sets which generate Boolean association rules. The algorithm uses an iterative method called layer search to generate (k + l) item sets from the k item sets. In this paper we describe a new algorithm which provides the way for data mining or data mining association on cloud environment so that we can achieve a better way to handle a large amount of data.

In this paper we discuss several technical issues related to security concern. We provide here an overview of executing data mining services on grid. The rest of this paper is arranged as follows: Section II introduces Cloud Computing; Section III describes about data mining; Section IV shows the Recent Scenario; Section V describes the proposed method. Section VI describes Conclusion and future prospect.

II. CLOUD COMPUTING

Cloud computing is the delivery of computing as a service rather than a product, whereby shared resources, software, and information are provided to computers and other devices as a utility (like the electricity grid) over a network (typically the Internet). A Cloud is a type of parallel and distributed system consisting of a collection of interconnected and virtualized computers that are dynamically provisioned and presented as one or more unified computing resources based on service-level agreements established through negotiation between the service provider and consumers.

The increased degree of connectivity and the increasing amount of data has led many providers and in particular data centers to employ larger infrastructures with dynamic load and access balancing.

There are several reasons to adopt cloud computing like
cost, scalability, business agility, and disaster recovery. Cloud computing is a model for enabling convenient, on-demand network access to a shared pool of configurable computing resources that can be rapidly provisioned and released with minimal management effort or service provider interaction.

This cloud model promotes availability and is composed of:
1) Five essential characteristics:
   a) On-demand self-service
   b) Broad network access
   c) Resource pooling
   d) Rapid elasticity
   e) Measured Service
2) Three service models:
   a) Cloud Software as a Service (SaaS)
   b) Cloud Platform as a Service (PaaS)
   c) Cloud Infrastructure as a Service (IaaS)
3) Four deployment models:
   a) Private cloud
   b) Community cloud
   c) Public cloud
   d) Hybrid cloud
4) Key enabling technologies include:
   a) Fast wide-area networks
   b) Powerful, inexpensive server computers
   c) High-performance virtualization for commodity hardware

In general, a public (external) cloud is an environment that exists outside a company’s firewall. It can be a service offered by a third-party vendor. It could also be referred to as a shared or multi-tenanted, virtualized infrastructure managed by means of a self-service portal.

A private (Internal) cloud reproduces the delivery models of a public cloud and does so behind a firewall for the exclusive benefit of an organization and its customers. The self-service administration interface is still in place while the IT infrastructure resources being collected are internal. In a hybrid cloud environment, external services are leveraged to extend or supplement an internal cloud.

III. DATA MINING TECHNIQUES
Here are several major data mining techniques have been developed and used in data mining projects recently including association, classification, clustering, prediction and sequential patterns. We will briefly examine those data mining techniques with example to have a good overview of them.

A. Association

Association is one of the best known data mining technique. In association, a pattern is discovered based on a relationship of a particular item on other items in the same transaction. For example, the association technique is used in market basket analysis to identify what products that customers frequently purchase together. Based on this data businesses can have corresponding marketing campaign to sell more products to make more profit.

B. Classification

Classification is a classic data mining technique based on machine learning. Basically classification is used to classify each item in a set of data into one of predefined set of classes or groups. Classification method makes use of mathematical techniques such as decision trees, linear programming, neural network and statistics. In classification, we make the software that can learn how to classify the data items into groups. For example, we can apply classification in application that “given all past records of employees who left the company, predict which current employees are probably to leave in the future.” In this case, we divide the employee’s records into two groups that are “leave” and “stay”. And then we can ask our data mining software to classify the employees into each group.

C. Clustering

Clustering is a data mining technique that makes meaningful or useful cluster of objects that have similar characteristic using automatic technique. Different from classification, clustering technique also defines the classes and put objects in them, while in classification objects are assigned into predefined classes. To make the concept clearer, we can take library as an example. In a library, books have a wide range of topics available. The challenge is how to keep those books in a way that readers can take several books in a specific topic without hassle. By using clustering technique, we can keep books that have some kind of similarities in one cluster or one shelf and label it with a meaningful name. If readers want to grab books in a topic, he or she would only go to that shelf instead of looking the whole in the whole library.

D. Prediction

The prediction as it name implied is one of a data mining techniques that discovers relationship between independent variables and relationship between dependent and independent variables. For instance, prediction analysis technique can be used in sale to predict profit for the future if we consider sale is an independent variable, profit could be a dependent variable. Then based on the historical sale and profit data, we can draw a fitted regression curve that is used for profit prediction.

E. Sequential Patterns

Sequential patterns analysis in one of data mining
technique that seeks to discover similar patterns in data transaction over a business period. The uncover patterns are used for further business analysis to recognize relationships among data.

IV. RECENT SCENARIO

In 2010, Kwuu W. Lin et al. [4] 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. [5] 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. [6] 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. [7] propose 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. [8] presents a specific method of implementing k-means 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. [9] 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. [10] 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.

V. PROPOSED ALGORITHM (CDM)

In this section we describe our proposed methodology that is Cloud Data Mining (CDM). Our approach is easily understandable with the help of the flowchart (Fig. 2).

By the below flowchart we can easily understand the whole phenomena. In our approach we can enter in the cloud or non cloud environment. Then we can perform data mining technique to find frequent patterns and relevant associations. After the completion of data mining technique we can deduce the cost in cloud and non cloud environment. Our Proposed algorithm is shown below:

Algorithm: CDM

1. Create a new database acting as non cloud environment
2. Again create a new database acting as a cloud environment.
3. We can add file from non cloud or from the main database to the cloud environment when applying for the operation.
4. We can enter in the data mining environment either through the cloud environment or from the non cloud environment.
5. Read the database.
6. While (object.read()!=-1)
   
   [Start Reading]

   [compute the occurrences]

   For i=1 to n iterations do
   
   Itemset [i]=count;
   Count++;

7. Make different associations based on the conditions.
8. Enter the minimum support.
   [Enter the minimum support]

   If (min-sup == msdb)
   
   [Prune(db, key)]
   
   Else
   
   [Empty]

   Prune(db, key)

   enter the min-sup
VI. CONCLUSION AND FUTURE PROSPECT

In this paper we want to generalize the formulation of data mining techniques with cloud computing environment. In data mining we want to find useful patterns with different methodology. The main issue with data mining techniques is that the space required for the item set and there operations are very huge. If we combine data mining techniques with cloud computing environment, then we can rent the space from the cloud providers on demand. This solution can solve the problem of huge space and we can apply data mining techniques without taking any consideration of space. This paper basically survey and analyze the utility for solving the above situation.

In future we concentrate on the real time scenario with their implementation.

REFERENCES


Dhresh Soni was born in Satna M.P India on August 29, 1984. He has done an M.E degree in computer science from Patel College of Science & Technology in 2012 Bhopal India, B.E degree in information technology in 2007 from RKDF Institute Bhopal India and diploma in information technology in 2004 from Dhar Poly Technique India.

Currently he is working as Assistant Professor in Sagar Institute Research Technology & Science Bhopal (M.P). He has worked as a game test engineer in global step pune and having work experience of 3 year 7 months. He has published 4 international papers.

Atish Mishra was born in M.P India on April 25, 1983. He has done an M.Tech degree in computer science from Madhav Institute of Technology and Science in 2009 Gwalior India, B.E. degree in computer science in 2005 from Madhav Institute of Technology and Science in 2009 Gwalior India. He is currently perusing the PhD in computer science from Madhav Institute of Technology and Science in 2009 Gwalior India.

Currently he is working as Associate Professor in Patel College of Science & Technology Indore (M.P). He has having work experience of 3 year 5 months. He has published 10 international papers.

Hitesh Gupta was born in M.P India on Oct. 19, 1982. He has done an M.E degree in computer science from Technocrat Institute of Science & Technology in 2010 Bhopal India, B.E. degree in information technology in 2005 from Bansal Institute Bhopal India.

Currently he is working as HOD- Associate Professor in Patel College of Science & Technology Bhopal (M.P). He has having work experience of 8 year 7 months. He has published 10 international papers.