A Frequent Pattern Mining Algorithm in Cloud Computing Environments using Association Rules Mining Algorithm

Avinash Sharma* and Dr.N.K.Tiwari**
*Research Scholar
avinashavi07@rediffmail.com
**Bansal Institute of Science and Technology, Bhopal

Abstract: Cloud computing is a new concept with a broad definition. Cloud computing is a new network computing paradigm based on IP architecture, and its potential lies in new business applications. 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. The main task associated with cloud computing is next generation data center transformation. In this paper we generalized the formulation of data mining technique with cloud computing environment and generate the result. 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 combine data mining techniques with cloud computing environment, then can rent the space from the cloud providers on demand. This solution can solve the problem of huge space with better application usage in low cost. We can apply data mining techniques without taking any consideration of space.

Keywords: Cloud computing, data mining, frequent pattern.

Introduction

Cloud Computing [1,2] is a new business model. 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 is a new concept with a broad definition. Cloud computing is a new network computing paradigm based on IP architecture, and its potential lies in new business applications. For the majority of operators and enterprises, the main task associated with cloud computing is next generation data center transformation. "Computing" generally refers to computing application; that is, any IT application in industry or in the market. Because network technologies are being converged, all applications in information, communication, and video are integrated on a unified platform. Likewise, computing in cloud computing refers to any integrated application. The key characteristic of cloud computing is not "computing" but "cloud." 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 adopt virtualization, service-oriented architecture, autonomic computing, and utility 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 over a network. This will ensure cloud computing becomes more widespread among enterprises, institutions, organizations, and operators. Cloud computing not only provides traditional IT resource usage and application services, but also supports full resource usage and application services such as IT, communications, video, mobile, and Internet of Things using a converged network infrastructure. Cloud computing technologies include Key feature of unified fabric, unified virtualization, and unified computing system. 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 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. 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 generateand-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. 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.

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. A cloud service has three distinct characteristics that differentiate it from traditional hosting. 1. It is sold on demand, typically by the minute or the hour. 2. It is elastic - a user can have as much or as little of a service as they want at any given time 3. Service is fully managed by the provider (the consumer needs nothing but a personal computer and Internet access). 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. Four deployment models:

1.1 Private cloud
1.2 Community cloud
1.3 Public cloud
1.4 Hybrid cloud

2. Three service models:

2.1 Cloud Software as a Service (SaaS)
2.2 Cloud Platform as a Service (PaaS)
2.3 Cloud Infrastructure as a Service (IaaS)

3. Five essential characteristics:

3.1 On-demand self-service
3.2 Broad network access
3.3 Resource pooling
3.4 Rapid elasticity
3.5 Measured Service

4. Key enabling technologies include:

4.1 Fast wide-area networks
4.2 Powerful, inexpensive server computers
4.3High-performance virtualization
4.4 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.

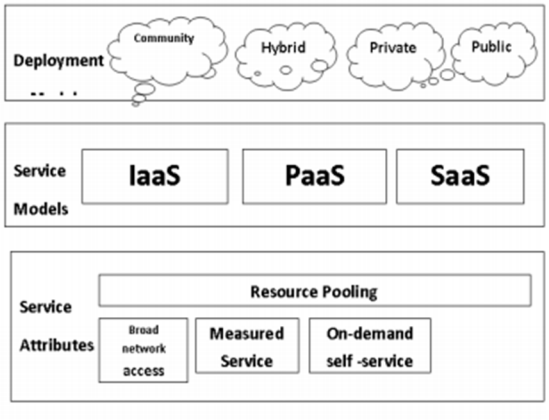

Fig.1 Cloud Computing

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." 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

Recent Scenario

In 2010, Kawuu 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] 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. [8] 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. [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.

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 cannot 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.

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 :

```
1) L(1)= find_frequent _1-itemsets(D);
2) For each transaction t belongs to D) 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);

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] A Weiss. “Computing in Clouds”, ACM Networker, 11(4):18-25,Dec.2007.
[2] R Buyya, CS Yeo, S Venugopal, Market-Oriented Cloud Computing:Vision, Hype, and Reality for Delivering IT Services as Computing Utilities. Proceedings of the 2008 10th IEEE International Conference on High Performance Computing and Communications. Vol.00, pp, 5-13, 2008.
[3] Shao Feng jing, Yu Zhong qing. Principle and Algorithm of Data Mining [M]. Beijing: China WaterPower Press, 2003. 2, 126-170.
[4] KawuuW.Lin , Yu-ChinLuo ,” Efficient Strategies for Many-task Frequent Pattern Mining in Cloud Computing Environments”,2010 IEEE.
[5] 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)
[6] Jiabin Deng, JuanLi Hu, Anthony Chak Ming LIU, Juebo Wu, “Research and Application of Cloud Storage”,2010 IEEE.
[7] Lingjuan Li , Min Zhang , “The Strategy of Mining Association Rule Based on Cloud Computing”, 2011 IEEE.
[8] T.R. Gopalakrishnan Nair, K.Lakshmi Madhuri , “DATA MINING USING HIERARCHICAL VIRTUAL K-MEANS APPROACH INTEGRATING DATA FRAGMENTS IN CLOUD COMPUTING ENVIRONMENT”,2011 IEEE.
[9] Lingjuan Li, Min Zhang, “The Strategy of Mining Association Rule Based on Cloud Computing”, 2011 International Conference on Business Computing and Global Informatization.
[10] Fabrizio Marozzo , Domenico Talia , Paolo Trunfio , “A Cloud Framework for Parameter Sweeping Data Mining Applications”, 2011 Third IEEE International Conference on Coud Computing Technology and Science.
[11] http://searchcloudcomputing.techtarget.com/ definition /cloud-computing
[12] http://www.makeuseof.com/tag/cloud-computing-work- technology-explained/
[13] C. Bohm, S. Berchtold, H. P. Kriegel, and U. Michel, "Multidimensional index structures in relational databases," in 1st International Conference on Data Warehousing and Knowledge Discovery (DaWak 99), Florence, Italy, 1999, pp.51-70.
[14] Dean, S. Ghemawat, and Usenix, "MapReduce: Simplified data processing on large clusters," in 6th Symposium on Operating Systems Design and Implementation (OSDI 04), San Francisco, CA, 2004, pp. 137-149.
[15] J. Han, J. Pei, and Y. Yin. Mining Frequent Patterns Without Candidate Generation. Proc. of ACM Int. Conf. on Management of Data (SIGMOD), 2000, pp. 1-12.
[16] [16] Gurudatt Kulkarni, Jayant Gambhir and Amruta Dongare, “Security in Cloud Computing”, International Journal of Computer Engineering & Technology (IJCET), Volume 3, Issue 1, 2012, pp. 258 - 265, ISSN Print: 0976 – 6367, ISSN Online: 0976 – 6375.
[17] Rohini G.Khalkar and Prof. Dr. S.H.Patil, “Data Integrity Proof Techniques in Cloud Storage”, International Journal of Computer Engineering & Technology (IJCET), Volume 4, Issue 2, 2013, pp. 454 - 458, ISSN Print: 0976 – 6367, ISSN Online: 0976 – 6375.