Applying Frequent Pattern Mining in Cloud Computing Environment

Dheresh Soni [1], Atish Mishra [2], Satyendra Singh Thakur [3], Nishant Chaurasia[4]
M.Tech Scholar (CSE), PCST Bhopal, India[1], Asst. Prof. Dept. of CSE, PCST Bhopal, India[2]
Asst. Prof. Dept. of CSE (HOD), PCST Bhopal, India[3], M.Tech Scholar (CSE) OIST Bhopal[4]

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.

Keywords

Cloud Computing, data mining, frequent pattern, cloud storage

1. 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.
Many approaches can be handled with high-dimensional and large-scale data, in which query processing is the bottleneck. “Algorithms for knowledge discovery tasks are often based on range searches or nearest neighbor search in multidimensional feature spaces” [3]. Business intelligence and data warehouses can hold a Terabyte or more of data. Cloud computing has emerged for the subsequently increasing demands of data mining. Map Reduce is a programming framework and an associated implementation designed for large data sets. The details of partitioning, scheduling, failure handling and communication are hidden by Map Reduce. Users simply define map functions to create intermediate <key, value> tuples, and then reduce functions to merge the tuples for special processing [4]. The basic conception of frequent pattern mining problem is to discover the pattern whose frequency of appearance in the database is greater than a specific threshold. An association rule is defmed as X=>Y, where X and Y are sets of items. The concept of association rule mining is to discover the sets of items tending to associate with the others in the database. The studies on association rule mining can be classified into two types, 1) the generate-and-test [w] (Apriori-like) approach and 2) the frequent pattern growth approach [5] (FP-growth-like).

The Apriori-like methods iteratively generate candidate itemset of size (k+ 1) from frequent itemset of size k and scan the database repetitively to test the frequency of each candidate itemset. Definitely, the Apriori-like methods suffer from the large number of candidate itemsets, especially when the support threshold is small. In view of this reason, Han et al. [5] proposed a novel data structure, named frequent pattern tree (FP-tree), in which the transactions are compressed and stored. A mining algorithm, namely FP-growth was also proposed for discovering the frequent patterns from the FP-tree. FP-growth needs only two scans on physical databases and therefore has a great improvement on the execution time.In this paper we discuss several technical issues related to security concern.We provide here an overview of executing data mining services on cloud. The rest of this paper is arranged as follows: Section 2 introduces Cloud Computing Section 3 describes about Data Mining Techniques; Section 4 shows the Recent Scenario; Section 5 describes the proposed methodology. Section 6 describes challenges in cloud computing. Section 7 describes about Conclusion and future prospect.

2. 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 as shown in Fig 1. 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. By distributing and replicating data across servers on demand, resource utilization has been significantly improved. Similarly web server hosts replicate images of relevant customers who requested a certain degree of accessibility across multiple servers and route requests according to traffic load. However, it was only when Amazon published these internal resources and their management mechanisms for use by customers that the term “cloud” was publicly associated with such elastic infrastructures – especially with “on demand” access to IT resources in mind. In the meantime, many providers have rebranded their infrastructures to “clouds”, even though this had little consequences on the way they provided their capabilities.

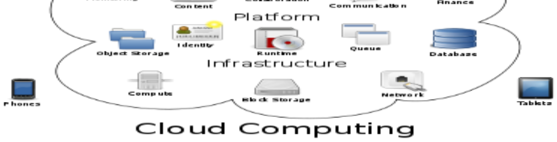

Fig 1. Cloud Computing Environment

A 'cloud' is an elastic execution environment of resources involving multiple stakeholders and providing a metered service at multiple granularities for a specified level of quality (of service). In other words, clouds as we understand them in the context of this document are primarily platforms that allow execution in various forms (see below) across multiple resources all of which have in common that they (directly or indirectly) enhance resources and services with additional capabilities related to manageability, elasticity and system platform independency.

3. 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.

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.

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.

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.

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.

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.

4. Recent Scenario

In 2010, Kawuu W.Lin et al. [6] 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. [7] 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. [8] 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. [9] 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. [10] 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.

5. Proposed Methodology

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. Data Mining (the analysis step of the Knowledge Discovery in Databases process, KDD), a relatively young and interdisciplinary field of computer science, is the process of discovering new patterns from large data sets involving methods from statistics and artificial intelligence but also database management. In contrast to for example machine learning, the emphasis lies on the discovery of previously unknown patterns as opposed to generalizing known patterns to new data. The goal of data mining is to discover the hidden useful information from large databases. Mining frequent patterns from transaction databases is an important problem in data mining field. As the size of database increases, the computation time and the required memory increase severely. Parallel and distributed computing techniques have attracted extensive attentions on the ability to manage and compute the significant amount of data in the past decades. The difficulty of mining large database launched the research of designing parallel and distributed algorithms to solve the problem. However, most of the past studies did not focus on the many-task issue that is very important, especially in cloud computing environments. In cloud computing environments, application is provided as service like Google search engine, meaning that it will be used by many users at the same time. In this paper, we propose 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. We want to design a cloud computing environment where the data sets are available on demand and the basis of the data set we apply further data mining techniques for finding the useful patterns. By the use of cloud computing we can utilize the space on demand, this is the advantage of cloud computing and we also apply data mining techniques.

6. Challenges in Cloud Computing

To that end, here's a rundown of ten key things both creators and users of cloud computing should continue to bear in mind.

1) Security

Cloud architectures don't automatically grant security compliance for the end-user data or apps on them, and so apps written for the cloud always have to be secure on their own terms. Some of the responsibility for this does fall to cloud vendors, but the lion's share of it is still in the lap of the application designer.

2) Complacency

A cloud computing-based solution shouldn't become just another passive utility like the phone system, where the owners simply puts a tollbooth on it and charges more and more while providing less and less. In short, don't give competitors a chance to do an end run around you because you've locked yourself into what seems like the best way to use the cloud, and given yourself no good exit strategy. Cloud computing is constantly evolving. Getting your solution in place simply means your process of monitoring and improving can now begin.

3) Client incomprehension

We're probably past the days when people thought clouds were just big server clusters, but that doesn't mean we're free of ignorance about the cloud moving forward. There are all too many misunderstandings about how public and private clouds (or conventional datacenters and cloud infrastructures) do and don't work together, misunderstandings about how easy it is to move from one kind of infrastructure to another, how virtualization and cloud computing do and don't overlap, and so on. A good way to combat this is to present customers with real-world examples of what's possible and why, so they can base their understanding on actual work that's been done and not just hypothetical where they're left to fill in the blanks themselves.

4) Preventing bottom-up adoption

Cloud infrastructures, like a lot of other IT innovations, don't always happen as top-down decrees. They may happen from the bottom up, in a back room somewhere, or on an employee's own time from his own PC.Examples of this abound: consider a New York Times staffer's experience with desktop cloud computing. Make a "sandbox" space within your organization for precisely this kind of experimentation, albeit with proper standards of conduct (e.g., not using live data that might be proprietary as a safety measure). You never know how it'll pay off.

5) Ad-hoc standards as the only real standards

The biggest example of this: Amazon EC2. As convenient as it is to develop for the cloud using EC2 as one of the most common types of deployments, it's also something to be cautious of. Ad-hoc standards are a two-edged sword.On the plus side, they bootstrap adoption: look how quickly a whole culture of cloud computing has sprung up around EC2. On the minus side, it means that much less space for innovators to create something open, to let things break away from the ad-hoc standards and can be adopted on their own. (Will the Kindle still be around in ten years?) Always be mindful of how the standards you're using now can be expanded or abandoned.

6) Over-utilization of capacity

Few things are more annoying to customers than promising something you can't deliver. The bad news is that in many industries, that's how things work: overbooking on airlines, for instance. Testing should always be standard practice. Robust, creative, think-out-of-the-box testing doubly so. Consider the way MySpace used 800 EC2 instances to test itself and see if they could meet anticipated demand for a new streaming music service. Their example involved using the cloud to test their native infrastructure, but there's no reason one couldn't use one cloud to generate test demand for another, and determine what your real needs are. And not just once, but again and again.

7) Under-utilization of capacity

This sort of thing's easier to deal with if you're the one buying the service, but what if you're the oneselling it? That's another reason why metrics and robust load testing are your best friends when creating cloud services. Also consider the possibility you're not selling enough kinds of services: is there room in your business plan for more granular, better-tiered service that might draw in a wider array of customers?

8) Network limitations

One word: IPv6. If you're deploying systems, using infrastructure or writing applications that aren't IPv6-aware now, you're building a time bomb under your chair. Think forward on every level, and encourage everyone building on top of your infrastructures to do the same thing.

9) Latency

Latency has always been an issue on the Internet; just ask your local World of War craft raiding guild. It's just as much of an issue in the cloud.Performance within the cloud doesn't mean much if it takes forever for the results of that performance to show up on the client. The latency that a cloud can introduce doesn't have to be deadly, and can be beaten back with both an intelligently planned infrastructure and smartly-written applications that understand where and how they're running.Also, cloud-based apps – and the capacity of cloud computing itself – are only going to be ramped up, not down, in the future. That means an arms race against increases in latency is in the offing as well. Just as the desktop PC's biggest bottlenecks are more often storage and memory, not CPU, the true source of cloud latency must be targeted and improved.

10) The Next Big Thing

The cloud isn't an endpoint in tech evolution; any more than the PC or the commodity server was final destinations. Something's going to come after the cloud, and may well eclipse it or render it redundant. The point isn't to speculate about what might come next, but rather to remain vigilant to change in the abstract. As the sages say, the only certainty is uncertainty, and the only constant thing is the next big thing.

7. 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

[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] 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.

[4] J. 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.

[5] 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.

[6] KawuuW.Lin,Yu-ChinLuo ," Efficient Strategies for Many-task Frequent Pattern Mining in Cloud Computing Environments",2010 IEEE.

[7] 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).

[8] Jiabin Deng, JuanLi Hu, Anthony Chak Ming LIU, Juebo Wu, "Research and Application of Cloud Storage",2010 IEEE.

[9] Lingjuan Li , Min Zhang , "The Strategy of Mining Association Rule Based on Cloud Computing", 2011 IEEE.

[10] T.R. Gopalakrishnan Nair, K.Lakshmi Madhuri , "Data Mining Using Hierarchical Virtual K-Means Approach Integrating Data Fragments In Cloud Computing Environment",2011 IEEE.