Review of Data Mining Techniques in Cloud Computing Database

Astha Pareek[1], Manish Gupta[2]
Research Scholar Department of CS & IT, the IIS University Jaipur[1]
Dy. Director (IT), RCEE, Dept. of Education, GoR, Jaipur, India[2]

Abstract

Data mining is the nontrivial extraction of implicit, previously unknown, and potentially useful information from data. It is the extraction of information from huge volume of data or set through the use of various data mining techniques. The data mining techniques like clustering, classification, neural network, genetic algorithms help in finding the hidden and previously unknown information from the database. Cloud Computing is a web-based technology whereby the resources are provided as shared services. The large volume of business data can be stored in Cloud Data centres with low cost. Both Data Mining techniques and Cloud Computing helps the business organizations to achieve maximized profit and cut costs in different possible ways. The main aim of the work is to implement data mining technique in cloud computing using Google App Engine and Cloud SQL.

Keywords

Data Mining, Google App Engine, Cloud SQL

1. Introduction

Data mining is the extraction of hidden information from the huge volume of data. The current business world is utilizing the data mining for gaining the insight into business strategies. There are no areas which are not affected by data mining .The clustering technique of data mining helps to segment the data according to the characteristic of the particular segment. This is helpful for detecting the loyal customer in the business world. The classification techniques of data mining help to classify the data on the basis of certain rules. This helps to frame policies for the future. The genetic algorithms help to find the best out of the given data. The data mining tools in the market provides an effective graphical user interface which helps the users to easily understand and analyze the data for strategic decision making.

Cloud Computing is a general term that refers to anything that "involves delivering hosted services over the Internet. Broadly it is characterized into three categories, namely: Software-as a Service (SaaS) ,Infrastructure-as-a- Service (IaaS) and Platform-as-a-Service (PaaS).Multi-Agent System is a problem solving "system composed of multiple interacting intelligent agents". Data Mining means "Discovery" of new knowledge that was not known before. It is "the analysis steps in the Knowledge Discovery and Databases process" (Nodine, Ngu, Cassandra and Bohrer, 2003).Data Warehouse refers to a data store which involves three stages, namely: staging, integration and access for reporting and analysis purposes.

Clustering is a technique by which similar records are grouped collectively. Usually this is done to give the end user a high level view of what is going on in the database. Clustering is sometimes used to denote segmentation - which most marketing people tell, is useful for coming up with a bird's eye vision of the business. K-means clustering is a data mining/machine learning algorithm used to cluster observations into groups of related observations without any prior knowledge of those relationships. The k- means algorithm is one of the simplest clustering techniques and it is commonly used in medical imaging, biometrics and related fields. Each object can be thought of as being represented by some feature vector in an n dimensional space, n being the number of all features used to describe the objects to cluster. The algorithm then randomly chooses k points in that vector space, these point serve as the initial centers of the clusters. Afterwards all objects are each assigned to center they are closest to. Usually the distance measure is chosen by the user and determined by the learning task. After that, for each cluster a new center is computed by averaging the feature vectors of all objects assigned to it. The process of assigning objects and recomposing centers is repeated until the process converges. The algorithm can be proven to converge after a finite number of iterations. The aim of K-means (or clustering) is this: We want to group the items into k clusters such that all items in same cluster are as similar to each other as possible. And items not in same cluster are as different as possible. We use the distance measures to calculate. Similarity and dissimilarity. One of the important concepts in K-means is that of centroid. Each cluster has a centroid. You can consider it as the point that is most representative of the cluster. Equivalently, centroid is point that is the "center" of a cluster.

Algorithm

1. Randomly choose k items and make them as initial centroids.
2. For each point, find the nearest centroid and assign the point to the cluster associated with the nearest centroid
3. Update the centroid of each cluster based on the items in that cluster. Typically, the new centroid will be the average of all points in the cluster.
4. Repeats steps 2 and 3, till no point switches clusters.

Data mining parameters include:

1. Association - Looking for patterns where one event is connected to another event.
2. Sequence or path analysis - Looking for patterns where one event leads to another later event
3. Classification - Looking for new patterns
4. Clustering - Finding and visually documenting groups of facts not previously known
5. Forecasting - Discovering patterns in data that can lead to reasonable predictions about the future, this area of data mining is known as predictive analytics.

2. Background and Related Work

By cloud we can say that it is an infrastructure that consists of services delivered through share data centers and appearing as a single point of access to consumers computing needs and also provides demanded resources or services over the internet. The concept of cloud computing does not provide facilities for the knowledge discovery and information retrieval. Furthermore, it is required that the so-called knowledge discovery should be in harmony with the structure, schema and architecture of that knowledge. The emerging knowledge cloud is considered insufficient to retrieve information effectively and thus, Chang, Yang and Luo (2011) undertook a study to an enormous number of ways simply because the items are so common that they cannot help but appear together. This is known as the rare item problem. It means that using the Apriori algorithm, we are unlikely to generate rules that may indicate rare events of potentially dramatic propose "an ontology-based agent generation framework for information retrieval in a flexible, transparent and easy way on cloud environment".

The ideal starting point is a data warehouse containing a combination of internal data tracking all customer contact coupled with external market data about competitor activity. Background information on potential customers also provides an excellent basis for prospecting. This warehouse can be implemented in a variety of relational database systems: Sybase, Oracle, Redbrick, and so on, and should be optimized for flexible and fast data access. An OLAP (On-Line Analytical Processing) server enables a more sophisticated end-user business model to be applied when navigating the data warehouse. The multidimensional structures allow the user to analyze the data as they want to view their business – summarizing by product line, region, and other key perspectives of their business. The Data Mining Server must be integrated with the data warehouse and the OLAP server to embed ROI-focused business analysis directly into this infrastructure. An advanced, to zero, RSAA takes a similar amount of time to that taken by Apriori generate low- support rules in amongst the high supports rules. Process-centric metadata template defines the data mining objectives for specific business issues like campaign management, prospecting, and promotion optimization. Integration with the data warehouse enables operational decisions to be directly implemented and tracked. As the warehouse grows with new decisions and results, the organization can continually mine the best practices and apply them to future decisions. This design represents a fundamental shift from conventional decision support systems. Rather than simply delivering data to the end user through query and reporting software, the Advanced Analysis Server applies users' business models directly to the warehouse and returns a proactive analysis of the most relevant information. These results enhance the metadata in the OLAP Server by providing a dynamic metadata layer that represents a distilled view of the data. Reporting, visualization, and other analysis tools can then be applied to plan future actions and confirm the impact of those plans.

As we stated in the introduction, the main objective of this work is to implement the K-Means algorithm in sets from the well-known real world data base "Machine learning repository, 2012".The first data set we used was "Iris Dataset" (Fisher,1936). It consists of 5 attributes and150 instances. The attributes are sepal width, sepal length, petal width, petal length and class label. It has three classes Iris flowers namely:

- Iris setos

- Iris versicolor
- Iris virginica

The following steps are performed to execute a data mining application through the Data Mining Cloud App

1) The user accesses the Website and submits his/her data mining application, by specifying: location of the input dataset, name of the data mining algorithm, and values of its parameters.
2) The Website inserts a set of tasks into the Task Queue on the basis of the data mining application submitted by the user .If the user submitted a single data mining application, one data mining task is inserted into the Task Queue. If the user submitted a parameter sweeping application, onetaskforeach combination of the input parameter values is created.
3) Each idle Worker picks a task from the Task Queue, and starts its execution on a virtual server.
4) Each Worker gets the input dataset from the location specified by the user To this end, a file transfer is performed from the Blob where the dataset is located to the local storage of the virtual server the Worker is running on.
5) After task completion, each Worker puts the result on a Blob.

3. Opportunities and Challenges

The use of the cloud provides a number of opportunities:

- It enables services to be used without any understanding of their infrastructure.
- Cloud computing works using economies of scale. It lowers the outlay expense for startup companies, as they would no longer need to buy their own software or servers. Cost would be by on-demand pricing. Vendors and Service providers claim costs by establishing an ongoing revenue stream.
- Data and services are stored remotely but accessible from ‘anywhere’.

In parallel there has been backlash against cloud computing:

- Use of cloud computing means dependence on others and that could possibly limit flexibility and innovation. The ‘others’ are likely become the bigger Internet companies like Google and IBM who may monopolise the market. Some argue that this use of supercomputers is a return to the time of mainframe computing that the PC was a reaction against.
- Security could prove to be a big issue. It is still unclear how safe outsourced data is and when using these services ownership of data is not always clear.
- There are also issues relating to policy and access. If your data is stored abroad whose FOI policy do you adhere to? What happens if the remote server goes down? How will you then access files? There have been cases of users being locked out of accounts and losing access to data.

4. Conclusion

It is widely accepted that K-Means algorithm is very popular clustering algorithm to analyze any real world problems. K-Means algorithm is more efficient algorithm for mining large Databases and Cloud computing provides solution for storing large database with less cost. So, in this paper, we focused the implementation of K-Means algorithm in the Cloud environment and the experimental results shows that it works well in the Cloud.

References

[1] W. J. Frawley, G. Piatetsky-shapiro, and C. J. Matheus. Knowledge discovery in databases: an overview, 1992.

[2] W.Wu and L. Gruenwald. Research issues in mining multiple data streams. In Proceedings of the First International Workshop on Novel Data Stream Pattern Mining Techniques, Stream KDD ’10, pages 56–60, New York, NY, USA, 2010.ACM.

[3] David E.Y. Sarna, Implementing And Developing Cloud Computing Applications, CRC Press https://cwiki.apache.org/MAHOUT/k-means-clustering.html.

[4] Weiss, A. (2007). Computing in the Clouds Networker.

[5] Wang, K., Xu, C. & Liu, B. (1999), Clustering transactions using large items, in ‘CIKM ’99: Proceedings of the Eighth International Conference on Information and Knowledge Management’, ACM Press, New York, NY, USA, pp. 483–490.

[6] Berson, Alex, Stephen Smith, and Kurt Thearling. Building data mining applications for CRM. New York: McGraw-Hill, 2000.

[7] Moving To The Cloud: Developing Apps in the

New World of Cloud Computing, By Dinkar Sitaram, Geetha Manjunath.
[8] The Cloud Computing Handbook - Everything You Need to Know about Cloud Computing, By Todd Arias.
[9] http://searchsqlserver.techtarget.com/definition/datamining.
[10] http://www.ijcaonline.org/volume15/number7/pxc387 2623.pdf.
[11] http://www.waset.org/journals/waset/v39/v39-72.pdf.
[12] http://www.estard.com/data_mining_marketing/data_mining_campaign.asp.
[13] http://dssresources.com/books/contents/berry97.html.
[14] http://www.marketingprofs.com/articles/2010/3567/the-nine-most-common-data-mining-techniques-usedin-predictive-analytics.
[15] http://www.thearling.com/.

Ms. Astha Pareek has started his career with software companies and worked for WIPRO Technology, Educomp solutions after completing post-graduation in computer science from Banasthali Vidyapeeth Jaipur. She is currently pursuing PH.D in Computer science &engineering from the IIS University Jaipur.

Dr. Manish Gupta, working in Education Department, Govt. of Rajasthan, Jaipur, India as the Dy. Director(Information Technology) handling the IT related projects like IT based House Hold Survey (online Child Tracking System), online GIS based school mapping with GPS coordinates of nearly 1.20 lacks schools across the state of of Rajasthan to create an huge database for planning purposes in the state under the national flagship program Sarva Shiksha Abhiyan(SSA)for universalization of elementary education. Data mining is the field of interest. In last 15 years, 28 research articles in the field of condensed matter physics, material science, electrodeposition in thin films and the cyclic voltametric analysis of different materials and data mining have been published in the Journals of National and International repute.