International Journal of Engineering & Technology, 7 (1.9) (2018) 265-267

International Journal of Engineering & Technology

Website: www.sciencepubco.com/index.php/IJET

Research paper

Data mining techniques and algorithms in cloud environment a review

K. Rajamani [1] *, D. Sheela [2]

[1] *Department of Computer Science and Engineering, New Prince Shri Bhavani College of Engineering and Technology*
[2] *Department of Electronics and Communication Engineering, Tagore Engineering College*
**Corresponding author E-mail: mithrankaruna@gmail.com*

Abstract

Cloud Computing is resourceful in which computing resources are made available on- demand to the user as needed. Data mining is a process of discovering interesting patterns from a large amount of data. The difficulty is in collecting these data and carrying out computations to get the significant information. Data mining techniques and applications can be effectively used in cloud computing environment. Data mining and the cloud computing are considered as major technologies. The data mining in cloud computing allows organizations to centralize the management of software and data storage. This paper provides a review of various data mining techniques and different types of algorithms in cloud computing which can be used for resource sharing.

Keywords: *Data Mining Techniques; Data Mining Algorithm; Cloud Computing.*

1. Introduction

1.1. Data mining

Data Mining is a logical process that is used to search through a large amount of data in order to find useful data [1] by using the following techniques. They are Clustering, Classification, Association, Regression, Attribute Importance, Anomaly Detection and Future extraction.

Data Mining is a part of the bigger framework, referred to as knowledge discovery in the database that covers a complex process from data preparation to knowledge modeling [2].

Components of Data Mining:

Association- Searching for patterns where the events are related to another event.

Sequence analysis – Searching for a pattern where one event leads to another event.

Classification – Searching for a new pattern.

Clustering – Searching a group of patterns which are not previously known.

Forecasting – Finding patterns of data that can lead to predict the future.

The Data Mining techniques are used to find the hidden and unknown information from the database [3].

Data Mining is a key step in knowledge discovery. It is the use of specific algorithms to extract patterns and knowledge from the data. The goal of data mining is to convert the large volume of data into useful information and knowledge [4].

1.2. Cloud computing

Cloud computing enables end users, small and medium-sized users to gain resources which are computational in nature. It also allows the users to use resources in the cloud and satisfy their necessities. Cloud user can access this data anytime anywhere in the world without any loss due to system failure. The cloud services are a platform as services (PaaS), Infrastructure as service (IaaS), and software as service (SaaS).

Infrastructure as a Service: IaaS is a service which is provided to the consumer for storage, processing, networks and the computing resources where the consumer can able to deploy and run the software which includes operating system and application.

Platform as a Service: PaaS is a service which is provided to the consumer where he can create an application using programming languages, services, and tools which are supported by the provider.

Software as a Service: SaaS is a service which is provided to the consumer who can use the application form the cloud infrastructure. The consumer does not have any worry about how to manage or control the cloud infrastructure [5].

The Cloud is Virtualized data center which can manage or control itself. Cloud is virtual computing resources that the cloud users can access it at anytime from anywhere.

The features of cloud computing:

1) It has a very large scale. Google cloud computing already has more than 100 million servers, Amazon, IBM, Microsoft, Yahoo and other "cloud" all has hundreds and thousands of servers.
2) The virtualization. The cloud computing allows users from any location, using a variety of terminal acquisition applications. Resources requested from the "cloud", rather than a fixed tangible entity.
3) High reliability. The multiple copies of data are maintained in cloud to avoid the loss of data and to increase the reliability. The cloud computing is more reliable than using the local computer.
4) The versatility. Cloud computing is not for a specific application, in the "cloud" can be constructed under the support of the ever-changing applications, with a "cloud" can support different applications running simultaneously.

5) The high scalability: "Cloud" size can be dynamically scalable to meet the needs of applications and user scale growth. [6].

2. Data mining in the cloud

Information mining systems and requisitions are sincerely needed in the distributed computing ideal model. As distributed computing is entering all the more in all degrees of business and experimental processing, it transforms into an incredible issue to be concerned by information mining. The Microsoft suite of cloud-based administrations presents another specialized sneak peak of Data Mining in the Cloud as "DMCloud". DMCloud permits you to perform some fundamental information mining assignments leveraging a cloud-based Analysis Services association. The data mining tasks include:
Analyze Key Influencers
Detect Categories
Fill From Example
Forecast
Highlight Exceptions
Scenario Analysis
Prediction Calculator
The Data mining is utilized within different requisitions, for example, medicinal services, personnel administration, math, science, in a different site. Data mining through cloud registering decreases the jumps that keep little organizations from profiting off the information mining instruments. We investigate how the information mining instruments like SaaS, PaaS, and IaaS are utilized within distributed computing to concentrate the data. Individuals utilize this characteristic to manufacture data posting and get data about distinctive themes via seeking in discussions and so forth. The organizations utilize this administration to see what sort of data is gliding on the planet-wide for their items or administrations and take activities dependent upon the information displayed. The data recovery commonsense model through the multi-executor framework with information mining in a distributed computing environment has been proposed. It is prescribed that clients might as well guarantee that the solicitation made to the IaaS is inside the extent of combined information warehouse and is clear and straightforward. The work for the multi-executor framework gets to be less demanding through the provision of the information mining calculations to recover serious data from the information warehouse [7].

3. Data mining algorithms

Various Data Mining algorithms and techniques are used for discovering the knowledge from the databases.

3.1. Classification

Classification is the data mining technique which is most commonly used, which employs a set of pre-classified examples to develop a model that can classify the population of records at large. The data classification process involves learning and classification. In Learning the training data are analyzed by the classification algorithm. In classification, test data are used to estimate the accuracy of the classification rules. If the accuracy is acceptable the rules then that can be applied to the new data tuples. [8]

3.2. Clustering

The unsupervised classification that is called as clustering or it is also known as exploratory data analysis in which there is no provision of labelled data. The main aim of clustering technique is to separate the unlabeled data set into a finite and discrete set of natural and hidden data structures. There is no provision of providing accurate characterization of unobserved samples that are generated from by same probability distribution
Broadly clustering has two areas based on which it can be categorized as follows:

- Hard clustering: In hard clustering same object can belong to a single cluster.
- Soft clustering: In this clustering same object can belong to different clusters.[9]

3.3. Regression

Regression technique can be used for prediction. Regression analysis can be used to relate the independent variables and dependent variables. Attributes are independent variables which are already known and response variables are what we are going to predict. Unfortunately, many real-world problems are not simply a prediction. For example, it is very difficult to predict if it depends on complex interactions of multiple predictor variables. Therefore, to forecast the future values more complex techniques (e.g., logistic regression, decision trees, or neural networks) are used. Neural networks can be used to create both classification and regression models. [10]

3.4. Association rule

Association and correlation are used to identify the frequently used items from the large data set. This type of technique helps businesses to make certain decisions, such as catalog design, cross-marketing, and customer shopping behavior analysis [11]. The main purpose is to discover rules linked with regularly co-occurring items, used for market basket analysis, root-cause analysis and cross-sell. The reason is to produce the precious information which describes links between data items from a large volume of data.

3.5. Neural networks

The neural network is a set of connected input/output units and each connection has a weight present with it. At learning phase, It can be able to predict the correct class labels of the input tuples by adjusting weights. Neural networks are used to derive the meaning from complicated data and it can be used to extract patterns. These are well suited for continuous-valued inputs and outputs. Neural networks are one of the best data mining techniques which are used for identifying patterns or trends in large set of data and it is very suitable for prediction or forecasting needs. [12]

3.6. CURE algorithm

CURE is an agglomerative type hierarchical clustering algorithm that uses portioning of the dataset. To handle a large database, a combination of partitioning and random sampling is used. In this algorithm, each partition is partially clustered which is first partitioned from the drawn sample of datasets. Partial clusters are then again clustered to obtain desired clusters [13].
Algorithm

1) Start considering each of the inputs as a separate cluster and each successive step combine the nearest pair of clusters.
2) C representative points are stored to calculate the distance between a pair of the cluster.
3) These are determined by first selecting C well-scattered points within the cluster and then diminishing them towards the center of the cluster by a fraction α.
4) Representative points of the cluster are used to calculate its distance from other clusters. [14]

3.7. BIRCH (balanced iterative reducing and clustering using hierarchies)

BIRCH algorithm is a type of hierarchical clustering algorithm. It is basically used for particularly very large databases because it reduces the number of input/output operations. Clustering is a one of the data mining algorithm which is used to group the similar objects so that the data can be identified easily. A cluster is, therefore, a collection of objects which are coherent internally, but clearly dissimilar to the objects belonging to other clusters. [15]

3.8. K- means clustering algorithm

It is one of the easiest unsupervised learning algorithms that are quite efficient in solving complex clustering problems. The procedure employs an easy and simple way to categorize a given data set into a certain number of clusters. [16]

K-means clustering algorithm is used to group the various observations which are related to each other without having any idea about the relationships existing among them. Some feature vectors in an n-dimensional space can be used to represent the objects, where n means the total number of the features that are being used for the description of the clusters. [17] Once it is done, the algorithm will choose k-points in the vector space. These k-points become initial centers of the cluster. Then all objects will be assigned to the center points which are at a minimum distance from them. This process will be repeated till the converging of the process.

4. Conclusion

Cloud Computing allows the users to retrieve information virtually at anytime from anywhere. The data mining in cloud computing allows the organization to centralize with the assurance of efficient, reliable and secure services for their users. Searching for frequent patterns in the database is one of the most important data mining problems. Here we explore how data mining algorithms and techniques are used in cloud computing to extract the information. The main issue with data mining technology is that the space required for the item set and their operations are very large. If we combine the data mining techniques with cloud computing environment, then we can rent the space from the cloud providers on demand.

References

[1] Mrs.Bharathi M.Rameageri,"Data mining and applications", Indian Journal of Computer Science and Engineering, Vol.1 No. 4, 2011.

[2] Dr. SankarRajagopal, "Customer Data Clustering using Data Mining Technique", International Journal of Databse Management Systems, Vol 3. N0 4, Nov 2011

[3] AsthaPareek, ManishGupta,"Reviewnof Data Mining Techniques in Cloud Computing Database", International Journal of Adavanced Computer Research, vol 2, no 2, june 2012.

[4] PeterMell, Timothy Grance: The NIST Definition of Cloud Computing, U. S. Department of Commerce, Special Publication 800-145.

[5] D. Taliaand P.Trunfio, How distributed data mining tasks can thrive as knowledge services Communications of the ACM. 53(2010)132-137. https://doi.org/10.1145/1785414.1785451.

[6] M. S. Chen, J. Han, and P. S. Yu. Data mining: an overview from database perspective. IEEE Trans. On Knowledge and Data Engineering, 5(1):866—883, Dec.199.

[7] Anuja R. Yeole, PoonamBorkar,"Survey Paper on Data Mining in Cloud Computing", International Journal of Science and Research, vol-4, issue-3, march 2015.

[8] Pavel Berkhin, Accrue Software, 1045 Forest Knoll Dr., SanJose, CA, 95129: Survey of Clustering Data Mining Technique.

[9] Prof. V. B. Nikam, VikiPatil: Study of Data Mining algorithm in cloud computing using Map Reduce Framework Journal of Engineering, Computers &Applied Sciences (JEC&AS) Volume 2, No.7, July2013.

[10] Jeffrey Voasand JiaZhang, ―Cloud Computing: New Wine or Just a New Bottle? ‖, Database Systems Journal vol. III, no. 3/2012 71IEEEInternet Computing Magazine.

[11] Yudho Giri Sucahyo, Ph. D, CISA: Introduction to Data Mining and Business Intellegence.

[12] Mansigera, Shivanigoel," Data Mining- Techniques, methods and Algorithms: A review on tools and their Validity", International Journal of Computer Application, Vol 113, No. 18, March 2015.

[13] TapasKanungo, Nathan S. Netanyahu, AngelaY. Wu: An Effi-cient k-Means Clustering Algorithm: Analysis and Implementation, ieee transactions on pattern analysis and ma-chine intelligence, vol. 24, no. 7, july2002.

[14] Prof. V. B. Nikam, VikiPatil: Study of Data Mining algorithm in cloud computing using Map Reduce Framework.

[15] XiaGeng, ZhiYang: Data Mining in Cloud Computing, International Conference on Information Science and Computer Applications (ISCA 2013).

[16] PeterMell, Timothy Grance: The NIST Definition of Cloud Computing, U. S. Department of Commerce, Special Publication 800-145.

[17] D. Taliaand P. Trunfio, How distributed data mining tasks can thrive as knowledge services Communications of the ACM. 53(2010)132-137. https://doi.org/10.1145/1785414.1785451.