International Journal of Scientific Engineering and Applied Science (IJSEAS) - Volume-1, Issue-8,November 2015
ISSN: 2395-3470
www.ijseas.com

SURVEY OF DATA MINING TECHNIQUES IN CLOUD COMPUTING

R.Kabilan[1], Dr.N.Jayaveeran[2]

[1]Research Scholar, P.G and Research Department of Computer Science, Khadir Mohideen College, Adirampattinam.

[2]Head, P.G and Research Department of Computer Science, Khadir Mohideen College, Adirampattinam.

Data Mining is a process of extracting potentially useful information from raw data. 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

This paper provides a survey of various data mining techniques are used in cloud computing. These techniques include Classification, Regression, Time Series Analysis, Prediction, Clustering, Summarization, Association Rules and Sequence Discovery.

1. DATA MINING TECHNIQUES

Data mining is highly helpful to collect relevant information from various sources of data. So it is highly helpful to achieve particular task. Data mining effort is normally used to create a descriptive mining model or a predictive mining model. Descriptive mining model is generally used to characterize the general properties of the data in the database. Predictive mining model is performing inference on the current data in order to make predictions.[7] The aim of predictive and descriptive model can be achieved by using a variety of data mining techniques as shown in figure1.[8] In order to find the characteristics of object evolution these techniques can be used. In decision making and to predict the future trends and behavior such information can be used.

Data Mining	Predictive Mining Model	Classification
		Prediction
		Regression Analysis
		Time Series Analysis
	Descriptive Mining Model	Association Rule
		Clustering
		Summarization
		Sequence Discovery

Figure 1. Data Mining Models

1.1 Classification: Classification based on categorical. This technique based on the supervised learning. It can be classifying the data based on the training set and values. These goals are achieve using a decision tree, neural network and classification rule (IF- THEN).

1.2 Prediction: Prediction models continuous-valued functions. It is used to predict missing or unavailable numerical data values rather than class labels. It is one of a data mining techniques that determine the relationship between independent variables and the relationship between dependent and independent variables [4].

1.3 Regression Analysis: Regression Analysis is a statistical method.[19] It can be used for numeric prediction and also used to model the relationship between one or more independent or predictor variables and a dependent or response variable.

1.4 Time Series Analysis: Sequences of values or events changing with time, typically measured at equal time intervals. Time series analysis is the process of using statistical techniques to model and make clear a time-dependent series of data points. Time series forecasting is a method of using a model to generate predictions (forecasts) for future events based on known past events. [9]

1.5 Association Rule: Association rule mining consists of first finding frequent item sets from which strong association rules in the form of A⇨B are generated. These rules also satisfy a minimum confidence threshold. Associations can be further analyzed to uncover correlation rules, which convey statistical correlations between item sets A and B.

1.6 Clustering: Clustering is a collection of similar data object. Dissimilar object is another cluster. Finding similarities between data according to the characteristics found in the data and grouping similar data objects into clusters. This technique based on the unsupervised learning (i.e. desired output for a given input is not known).

1.7 Summarization: Summarization is an abstraction of data for setting relevant task and giving an overview of data. It includes various summarizations like: mean, weighted mean, median, and mode for measuring the central tendency of data, and range, quartiles, inter quartile range, variance, and standard deviation for measuring the dispersion of data.

1.8 Sequence Discovery: It uncovers relationship among data [8]. It is a set of object in which each one is associated with its own timeline of events. Discovery of sequential pattern means finding sequential associations among items in the sequence database.

2. CLOUD COMPUTING

Cloud computing is an internet-based computing in which shared resource, software and information are supplied to computers and other devices on demand. Generally it is divided into three service models. They are: SaaS, PaaS, IaaS

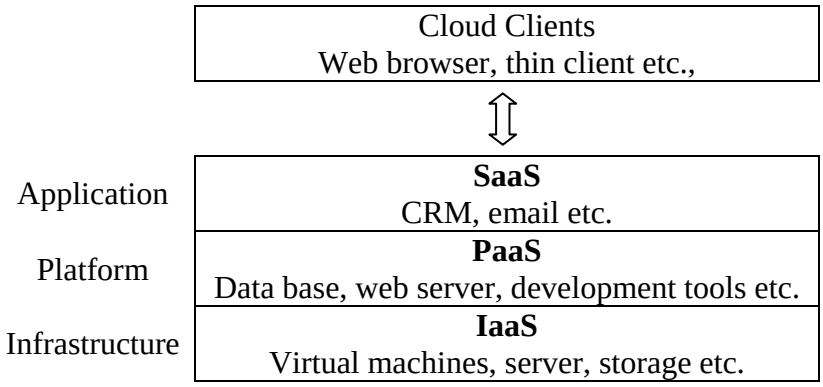

Figure 2. Cloud Computing

2.1 IaaS

Infrastructure-as-a-Service (IaaS) presents computer infrastructure, typically a platform virtualization environment, as a service. Rather than purchasing software, data center space, servers, network equipment, clients instead buy those resources as a fully outsourced service.

2.2 PaaS

Platform-as-a-Service (PaaS) offer a suitable situation for application developers can develop their own applications. In the PaaS models, cloud providers deliver a computing platform, typically including operating system, programming-language execution environment, database, and web server. Application developers can develop and run their software solutions on a cloud platform without any expenses and intricacy of buying and managing the underlying hardware and software layers.

2.3 SaaS

Software-as a Service **(**SaaS) is a model of software deployment where an application is hosted as a service provided to customers across the Internet. The users are not having any control to manage or manage the underlying cloud infrastructure including network, servers, operating systems, storage, or even individual application capabilities, with the possible exception of limited user specific application configuration settings.

3. DATA MINING IN CLOUD COMPUTING

Data mining techniques and applications are more and more essential in the cloud computing pattern. As cloud computing is penetrating in all type of computing, it is an essential area to be focused by data mining. Data mining in cloud computing is one of the methods of gathering structured information from unstructured or semi-structured web based data sources.

Using data mining through Cloud Computing minimize the barriers that keep small companies from benefiting of the data mining instruments. 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 permit the users to get back significant

information from virtually integrated data warehouse that reduces the cost of infrastructure and storage.

CDM (Cloud Data Mining) offers fabulous potential for analyzing and extracting the information in various fields of human activities: business, economics, health care medicines, heredity, biology, pharmacy, advertising, etc. The use of this technology should allow that with just a few clicks of the mouse one can reach the preferred information about clients, behavior, wellbeing, purchasing power, and regularity of purchases of certain items.

Cloud provides tools that can "handle" large volume of data, which cannot be processed efficiently and at reasonable cost using certain techniques.

4. DATA MINING TECHNIQUES IN CLOUD COMPUTING.

4.1 Time Series Analysis

Virtualized Cloud platforms and online services make a very large volume of monitored data in the form of time series; it becomes very difficult to process this complex data by conservative approaches. An approach on a production of Cloud encompassing IaaS and PaaS service models and experimental results [10] validate efficient and effective in capturing the metrics causing performance anomalies in large time series of datasets.

4.2 Association Rule

An association rule mining helps in finding relation between the items or item sets in the given data. The association rule mining algorithm in Hadoop [11] evaluates testing it in the cloud (EC2) by increasing the number of nodes in the testing set up. The input data is divided among the nodes. Further, the data transfer among the nodes and the situations like a storage node dies or, what would happen if some nodes in the cluster does not run, are taken care by Hadoop. This adds a great deal of robustness and scalability to the system.

Apriori algorithm is a classic data mining algorithm of association rules of data. Apriori algorithm is used in the cloud computing environment to solve traditional problems encountered in the traditional Apriori data mining. Apriori Algorithm Research [5] achieved high extension capability based on Hadoop platform, and proves the possibility of association rule mining algorithm and cloud computing technologies.

There are numerous data in the cloud database and among these data, much potential and valuable knowledge are implicit. The key point is to discover and pick up the useful knowledge automatically. An association rule [6] is one of the main models in mining out these data, and it mainly focuses on the relationship among different areas in the data.

Cloud computing is an effective platform for data mining. To gain more experience in cloud-assisted data mining, the association rule based algorithm, Apriori [17] proved how data

mining algorithms can be adjusted to fit the increasing demand for parallel computing environment of cloud.

4.3 Classification, Regression, Summarization

The applications of cloud computing are practically boundless. With the right middleware, a cloud computing system could execute all the programs a normal computer could run. Potentially, everything from generic word processing software to customized computer programs designed for a specific company could work on a cloud computing system. We can use cloud computing for fingerprint data storage and recognition. There are several areas we can visualize, where data mining can be applied. A particular data mining algorithm is usually an instantiation of the model-preference-search components. The models that can be useful for fingerprint data handling are: Classification, Regression, Clustering and Summarization. [12]

Cloud computing supplies cheap and efficient remedies for storing and analyzing mass data .The strategy of classification in cloud computing environment [18] can records higher efficiency.

4.4 Clustering

Clustering model for assessing SaaS will help to evaluate possible software services on the cloud computing by using Data Mining Clustering algorithms. The clustering model [13] would be highly useful to software service providers to evaluate their own services to the cloud users. It helps service provider to increase availability of software services on the cloud computing environment suitable for cloud users needs. It also helps cloud users to evaluate potential software services available on the cloud computing environment.

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 [15] for storing large database without any cost.

4.5 Prediction

Prediction framework[15] to predict critical events in virtualized cloud computing environment. This framework combines features of hash table and reversal pattern tree structure to increase the prediction speed for real time critical event pattern recognition and prediction.

The real benefit of Cloud Computing is the ability of a Cloud Client to be energetically scalable based on its use. This has great implication on cost saving as resources are not paid for when they are not used. Dynamic scalability is achieved through virtualization. The downside of virtualization is that they have a non-zero setup time that has to be considered for an efficient use of the platform. It follows that a prediction method would greatly aid a Cloud

Client in making its auto-scaling decisions. In resource usage prediction algorithm [16] uses a set of historic data to recognize similar usage patterns to a current window of records that occurred in the past. The algorithm then predicts the system usage by interpolating what follows after the identified patterns from the historical data.

5. CONCLUSION

Data mining technologies provided through Cloud computing is an essential characteristic for present day businesses to make proactive, knowledge driven decisions, as it helps them have future trends and behaviors predicted. This paper provides an survey of data mining techniques like Classification, Regression, Time Series Analysis, Prediction, Clustering, Summarization, Association Rules and Sequence Discovery.

6. REFERENCES

[1] Smita, Priti Sharma "Use of Data Mining in Various Field: A Survey Paper" IOSR Journal of Computer Engineering (IOSR-JCE) e-ISSN: 2278-0661, p- ISSN: 2278-8727Volume 16, Issue 3, Ver. V (May-Jun. 2014), PP 18-21

[2] Bhagyashree Ambulkar, Vaishali Borkar Data Mining in Cloud Computing. MPGI National Multi Conference 2012 (MPGINMC-2012) 7-8 April, 2012 "Recent Trends in Computing" Proceedings published by International Journal of Computer Applications® (IJCA)ISSN: 0975 – 8887

[3] CH.Sekhar, S Reshma Anjum" Cloud Data Mining based on Association Rule" CH.Sekhar et al, / (IJCSIT) International Journal of Computer Science and Information Technologies, Vol. 5 (2), 2014, 2091-2094

[4] Aakanksha Bhatnagar, Shweta P. Jadye, Madan Mohan Nagar "Data Mining Techniques & Distinct Applications: A Literature Review" International Journal of Engineering Research & Technology (IJERT) Vol. 1 Issue 9, November- 2012

[5] Zhang Danping Yu Haoran and Zheng Linyu, Apriori Algorithm Research Based on Map-Reduce in Cloud Computing Environments. The Open Automation and Control Systems Journal, 2014, 6, 368-373

[6] Zhu Tianxiang, Sun Shuhui, Zhang Dan, Liu Xin , Data Mining of the Association Rules Based on the Cloud Database. Proceedings of the 2nd International Conference on Computer Science and Electronics Engineering (ICCSEE 2013)

[7] Nikita Jain, Vishal Srivastava "DATA MINING TECHNIQUES: A SURVEY PAPER" IJRET: International Journal of Research in Engineering and Technology, Volume: 02 Issue: 11 | Nov-2013,

[8] Dr. M.H.Dunham,"Data Mining, Introductory and Advanced Topics", Prentice Hall, 2002.

[9] Time Series Analysis and Forecasting with Weka
http://wiki.pentaho.com/display/DATAMINING/

[10] Jehangiri, A.I , Yahyapour, R. ; Wieder, P. ; Yaqub, E. ; Kuan Lu "Diagnosing Cloud Performance Anomalies Using Large Time Series Dataset Analysis" Published in:Cloud Computing (CLOUD), 2014 IEEE 7th International Conference

[11] Pallavi Roy, MINING ASSOCIATION RULES IN CLOUD, Research paper ,August 2012.

[12] Kalyani Mali, Samayita "Bhattacharya Fingerprint Database Handling Using Cloud Computing With Added Data Mining and Soft Computing Features"
International Journal of Emerging Technology and Advanced Engineering Website: www.ijetae.com (ISSN 2250-2459, ISO 9001:2008 Certified Journal, Volume 3, Issue 2, February 2013) 610 .

[13] Mrs. Dhanamma Jagli, Mrs. Akanksha Gupta Clustering Model for Evaluating SaaS on the Cloud International Journal of Application or Innovation in Engineering & Management (IJAIEM) Volume 2, Issue 12, December 2013.

[14] Yuanyao Liu," Wu Critical System Event Prediction in Virtualized Cloud Computing Systems"

[15] A. Mahendiran, N. Saravanan, N. Venkata Subramanian and, N. Sairam, Implementation of K-Means Clustering in Cloud Computing Environment Research Journal of Applied Sciences, Engineering and Technology 4(10): 1391-1394, 2012 ISSN: 2040-7467

[16] Eddy Caron, Frederic Desprez and Adrian Muresan "Forecasting for Cloud computing on-demand resources based on pattern matching" inria-00460393, version 1 - 9 Mar 2010

[17] Juan Li Pallavi Roy Samee U. Khan Lizhe Wang Yan Bai "Data Mining Using Clouds: An Experimental Implementation of Apriori over MapReduce"

[18] Lijuan Zhang, Shuguang Zhao "The Strategy of Classification Mining Based on Cloud Computing" International Workshop on Cloud Computing and Information Security (CCIS 2013)

[19] Jiawei Han and Micheline Kamber "Data Mining concepts and Techniques".