http://ansinet.com/itj

ITJ

ISSN 1812-5638

INFORMATION TECHNOLOGY JOURNAL

ANSInet
Asian Network for Scientific Information
308 Lasani Town, Sargodha Road, Faisalabad - Pakistan

Information Technology Journal 12 (18): 4637-4641, 2013
ISSN 1812-5638 / DOI: 10.3923/itj.2013.4637.4641

A Dynamic Clustering Algorithm for Cloud Computing

[1,2]Zhongxue Yang, [1]Xiaolin Qin, [2]Wenrui Li and [3]Yingjie Yang
[1]College of Computer Science and Technology, Nanjing University of Aeronautics and Astronautics, 210016, Nanjing, China
[2]School of Mathematic and information Technology, Nanjing Xiaozhuang University, 211171, Nanjing, China
[3]Centre for Computational Intelligence, De Montfort University, Leicester, LE19BH, UK

Abstract: A novel dynamic clustering algorithm for cloud computing is proposed in this study. Dynamic clustering algorithm originates from K-means algorithm, however, an important characteristic of dynamic clustering algorithm is dynamic in that the centroid of the new cluster is updated in each iteration and some certain points near the boundary may be labeled to a new cluster in next iteration. Cloud computing is a new generation of computation platform with the nature of widely distributed and heterogeneous environment. Hadoop as a cloud platform and Map/Reduce as a distributed computing architecture are described and K-means clustering algorithm is illustrated for comparison to evaluate the performance in this study as well. Following this, experiments on KDD DATA datasets are conducted and the result shows that Dynamic Clustering algorithm can exhibit an excellent performance with higher attack detection rate and lower false positive rate while comparison with K-means clustering algorithm.

Key words: Dynamic clustering, Map/Reduce, K-means, cloud computing

INTRODUCTION

Cloud Computing is a distributed platform providing on demand computing resources and rich information and enabling convenient and on demand accessing of the resources. Cloud computing offers a shared pool of resources that are available on customer's demand and can be accessed at anytime from anywhere. Cloud computing offers the prospect of facilitating the development of large scale, flexible computing infrastructures and provides an extensible and powerful environment for growing amounts of services and data by means of on-demand self-service (Bhupendra and Kapoor, 2013). More and more people or organizations, therefore, have outsourced their data to the cloud computing and cloud computing has had a great impact on their services and efficiency. At the same time, huge amount of data are more easier to be collected and stored in cloud storage. The great amount of data stored in cloud storage continues to grow fast which contains very fruitful valuable knowledge. Hence, such large databases have led to the emergence of a field called data mining and knowledge discovery in databases (Shindler *et al*., 2001). Data Mining refers to extracting or mining knowledge from large amounts of data and clustering typically groups data into sets in such a way that the intra-cluster similarity is maximized while the inter-cluster similarity is minimized. Clustering is usually associated with large scale data and often requires a large amount of computer resources. Cloud computing suits well in solving these large scale data or large amount resources, with its promise of provisioning virtually infinite resources. In the adaptation of typical clustering models for the clouds, it must be reduced to a distributed framework that can successfully discover the knowledge from cloud computing.

In this study, a novel dynamic clustering algorithm for cloud computing is proposed. The remainder of this study is organized as follows. In section 2, K-means clustering algorithm is overviewed. Section 3 introduces Hadoop and Map/Reduce, a kind of cloud computing platform and computation architecture. Dynamic clustering algorithm for cloud computing is proposed in section 4. Simulation experiment is given in 5th section and the performance of the Dynamic Clustering algorithm is evaluated as well. Finally, this study is concluded in section 6.

K-MEANS CLUSTERING ALGORITHM

The clustering problem is the ordering of a set of data into groups, based on one or more features of the data. Cluster analysis (Duran and Odell, 1974; Jain *et al*., 1999;

Corresponding Author: Zhongxue Yang, College of Computer Science and Technology, Nanjing University of Aeronautics and Astronautics, 210016, Nanjing, China

Kotsiantis and Pintelas, 2004) is an unsupervised learning method that constitutes a main role of an intelligent data analysis process. It is used for the exploration of inter-relationships among a collection of patterns, by organizing them into homogenous clusters. It is called unsupervised learning because unlike classification (known as supervised learning), no a priori labeling of some patterns is available to use in categorizing others and inferring the cluster structure of the whole data.

Most previous clustering algorithms focus on numerical data whose inherent geometric properties can be exploited naturally to define distance functions between data points. Moreover, much of the data existed in the databases is categorical, where attribute values cannot be naturally ordered as numerical values. Mainly Clustering is try to group the similar type objects into one cluster and a cluster is chosen in order to minimize some measure of dissimilarity. K-Means clustering is a well known partitioning method to classify data in which the given data set is divided into K number of clusters. The results of partitioning method are a set of K clusters, each object of data set belonging to one cluster. Each cluster is associated with a centre point called centroid and each point is assigned to a cluster with the closest centroid. The centroid is nothing but the mean of the points in the cluster. Euclidean Distance method is used to calculate the distance of different points.

Many researchers have performed such related works in cloud computing. Mahendiran *et al.* (2012) have implemented K-Means clustering algorithm in cloud computing environment. It is obtained that both Data Mining techniques and cloud computing helps the business organizations to achieve maximized profit and cut costs in different possible ways. Liu and Cheng (2012) have adapted K-means algorithm to cloud computing as well and this method can be used to handle massive data effectively in cloud environment which can overcome the processing bottleneck of traditional algorithm. In fact, K-means clustering algorithm which is one of the very popular and high performance clustering algorithms, is used in cloud. Malathy and Somasundaram (2012) proposed the Reservation Cluster approach for performance enhancement in cloud computing. The concept of reservation cluster is to schedule the unscheduled tasks. Unscheduled tasks are sent to the reservation cluster and in this cluster all the tasks are scheduled simultaneously without any iteration. It reduces the amount of computation time and resource usage and allows better performance. Chen and Qiao (2011) introduced K-means algorithm to Hadoop platform and presented the MapReduce. They combined the K-means with data mining technology to implement the effectiveness analysis and application of the cloud computing platform. Shindler *et al.* (2011) implemented a fast and accurate k-means clustering for large datasets. Obtained from their works, fast and accurate clustering algorithm seems to be a better option to adopt in clouds.

CLOUD COMPUTING PLATFORM

Hadoop is an open source architecture and is introduced as a distributed computing framework by Google. Hadoop, with the main features including strong expansion ability, low cost, high efficiency, good reliability, free open source and good portability, etc., also is a cloud computing platform that could easier develop and concurrent process large-scale data. The cluster of Hadoop is the typical Master/Slaves framework and it consists of two parts: HDFS (Hadoop Distributed File System) and MapReduce computing model (Marakas, 2003; Zhao *et al.*, 2001). HDFS is the cornerstone of distributed computation which adopts M/S framework. The operations that HDFS could perform include creating, deleting, moving or renaming, etc. And the framework is similar to traditional hierarchical filing system. Map/Reduce is a kind of high-efficiency distributed computational model. Its operating principle is that decomposing the task for hundreds of small tasks and sending to the computer cluster. Each computer then sends back the information of their parts. MapReduce is a programming model and a distributed computing framework. It was first developed by Google to process very large amounts of raw data that it has to deal with on a daily basis, like indexed Internet documents and web requests logs. With the MapReduce programming model, developer just need to specify two functions: Map and Reduce. The Map Function accepts a set of data and converts them into a key/value list<key, value>. MapReduce applications get a list of key-value pairs as an input. The Map method processes each key-value pair in the input list separately and outputs one or more key-value pairs as a result. MapReduce will rapidly integrate these feedbacks and form answers. Map/Reduce source is the Map and the Reduce function of functional programming.

Pair: The Reduce Function receives the list formed by the Map Function, then reduces the key/value list according to their keys and outputs new key/value list<key, value>. The process of Map/Reduce is shown in Fig. 1(Google Developer, 2012).

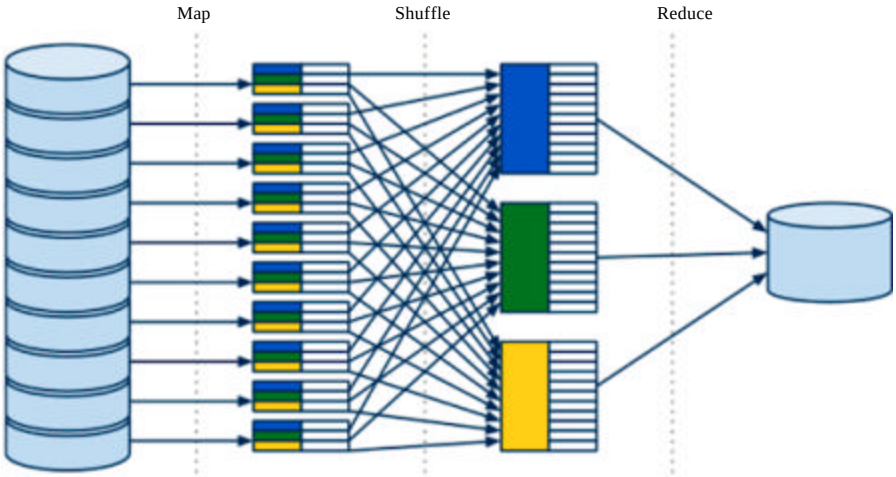

Fig. 1: Framework of MapReduce

DYNAMIC CLUSTERING ALGORITHM FOR CLOUD

Dynamic clustering algorithm looks like K-means clustering algorithm, but the centroid of each cluster may be updated in a new clustering procedure. Moreover, in the clustering procedure, a sample may be moved to a new cluster. Dynamic clustering algorithm works as below simply.

Step 1: In the dataset D has n samples totally, select k points according to the number of datacenters in D and k = n initially, so one sample is one cluster

Step 2: Calculate the k initial cluster center, i.e. each cluster centroid $g_1, g_2, \ldots, g_k$

Step 3: Calculate the distance from each sample x_i (i = 1, 2,…, n) to the centroid g_j (j = 1, 2,…, k) according to Euclidean Distance equation:

$$d_{ij} = \sqrt{\sum_{l=1}^{p}(x_{it} - x_{jt})} \quad (i, j = 1, 2, \cdots n)$$

and find the minimal distance. And then, $x_i \in G'_a$, where the centroid of new cluster G'_a is modified as g'_a and $\min |x_i - g_a| \equiv |x_i - g_a'|$

Step 4: Calculate the new centroid of each cluster $g'_1, g'_2, \ldots, g'_k$

Step 5: If $g_a \neq g'_a$, then go to step 3

Step 6: k = k-1 and repeat step2, until k = K, where K is the number of clusters desired

The Map/Reduce adaption of the Dynamic Clustering algorithms can be described as follows:

Step 1: Input data set (through a data file) is partitioned into N parts (controlled by the run-time system, not your program). Each part is sent to a map

Step 2: In the map function, the distance between each point and each cluster center is calculated and each point is labeled with the center index to which the distance is the smallest. Map outputs the key-value pairs of label assigned to each point and the coordinates of the point

Step 3: All data points of the same current cluster (based on the current cluster centers) are sent to a single reducer. In the reduce function, new cluster center coordinates are easily computed. The output of the reducer is consequently the cluster index and its new coordinates

Step 4: The new cluster coordinates are compared to the original ones. If the difference is within a preset threshold, then program terminates and we have found the clusters. If not, use the newly generated cluster centers and repeat step 2-4

Dynamic Clustering algorithm based on Cloud Computing (DCCC) makes an initial guess of the solution, in this case the clustering and at each following iteration it improves the accuracy of the solution. Also, it is not possible to reduce the whole algorithm to the MapReduce model. However, the content of a whole iteration can be reduced to the MapReduce model. Actually, in Map procedure, algorithm will find the closest centroid and assign the points to its cluster and the input is (cluster id, points), the output is (new cluster id, object). Similarly, the procedure of Reduce, system will try to find which object

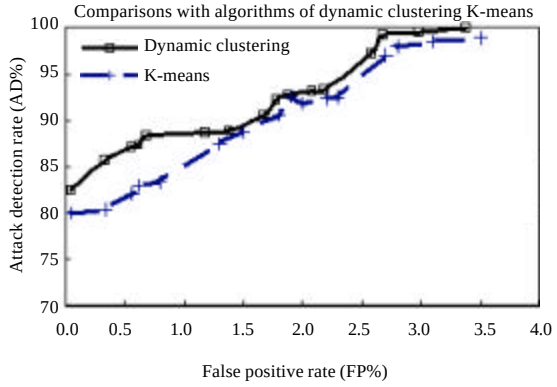

Fig. 2: Comparison of dynamic clustering and K-means algorithm

is the most central and assign it as a new centroid to the cluster and the input will be (cluster id, (list of all objects in the cluster)), (cluster id, new medoid) is the final outputs.

SIMULATIONS

KDD Cup (1999) is used to evaluate the algorithm and attack detection rate defines the percentage of the number of detected attacks in relation to the number of overall attacks. False positive rate, however, is defined as the percentage of the number of misclassified processes in relation to the number of normal processes. The detection algorithm with higher attack detection rate and lower false positive rate is regarded as good performance. For comparison, the same test data set are applied to evaluate Dynamic Clustering Algorithm and K-means algorithm, respectively. The results of experiment on anomaly detection with both algorithms are shown in Fig. 2.

The detection ROC curves for Dynamic Clustering and K-means algorithm in cloud computing are shown together in Fig. 2. Each curve depicts the relationships between attack detection rate and false positive rate based on different threshold. From the figure, it derives that in that providing the lowest false positive rate of 3.3% when the attack detection rate reaches 100% while K-means algorithm just providing 3.7 and 98.9%. Moreover, the convergence of Dynamic Clustering algorithm is more faster than K-means also as shown in Fig. 2.

CONCLUSION

In this study, a novel dynamic clustering algorithm for cloud computing is proposed. The centroid of the new cluster is updated in each iteration and some certain points near the boundary may be labeled to a new cluster in next iteration while using Dynamic Clustering algorithm. Hadoop as a cloud platform and Map/Reduce as a distributed computing architecture are described as well in this study. K-means clustering algorithm is illustrated for comparison to evaluate the performance. Following this, experiments on KDD DATA datasets are conducted and simulation result shows that Dynamic Clustering algorithm can exhibit an excellent performance with higher attack detection rate and lower false positive rate while comparison with K-means clustering algorithm.

ACKNOWLEDGMENT

The authors would like to thank the support by the National Natural Science Foundation of China under Grant No. 61202136 and the Natural Science Foundation of Jiangsu Education Department under Grand No. 09KJD520011 and No. 12KJD520008, respectively. The work is also supported by Jiangsu Province Scholarships of Overseas Studies.

REFERENCES

Bhupendra, P. and R.K. Kapoor, 2013. Dynamic VM allocation algorithm using clustering in cloud computing. Int. J. Adv. Res. Comput. Sci. Software Eng., 3: 143-150.

Chen, H. and Y. Qiao, 2011. Research of cloud computing based on the hadoop platform. Proceedings of the International Conference on Computational and Information Sciences, October 21-23, 2011, Chengdu, China, pp: 181-184.

Duran, B. and P. Odell, 1974. Cluster Analysis: A Survey. Springer-Verlag, New York.

Google Developer, 2012. MapReduce python overview. https://developers.google.com/appengine/docs/python/dataprocessing/

Jain, A.K., M.N. Murty and P.J. Flynn, 1999. Data clustering: A review. ACM Comput. Surveys, 31: 264-323.

KDD Cup, 1999. Data. Online. http://kdd.ics.uci.edu/

Kotsiantis, S.B. and P.E. Pintelas, 2004. Recent advances in clustering: A brief survey. WSEAS Trans. Inform. Sci. Appl., 1: 73-81.

Liu, S.P. and Y.L. Cheng, 2012. Research on K-Means algorithm based on cloud computing. Proceedings of the International Conference on Computer Science and Service System, August 11-13, 2012, Nanjing, pp: 1762-1765.

Mahendiran, A., N. Saravanan, N.V. Subramanian and N. Sairam, 2012. Implementation of K-means clustering in cloud computing environment. Res. J. Applied Sci. Eng. Technol., 4: 1391-1394.

Malathy, G. and R. Somasundaram, 2012. Performance enhancement in cloud computing using reservation cluster. Eur. J. Scient. Res., 86: 394-401.

Marakas, G., 2003. Modern Data Warehousing, Mining and Visualization: Core Concepts. Prentice Hall, Upper Saddle River, New Jersey.

Shindler, M., A. Wong and A. Meyerson, 2011. Fast and accurate k-means for large datasets. NIPS 2011: Granada, Spain, pp: 2375-2383.

Shindler, M., A. Wong and A.J. Han *et al*., 2001. Data Mining: Concepts and Techniques. Morgan Kaufmann, San Francisco.

Zhao, W., H. Ma, Y. Fu and Z. Shi, 2001. Research on parallel K means algorithm design based on hadoop platform. Comput. Sci., 38: 166-176.