IJCSE International Journal of Computer Sciences and Engineering Open Access

Review Paper Volume-4, Issue-8 E-ISSN: 2347-2693

A Novel Algorithm for Big Data Clustering

Vishal Kumar Gujare[1], Prof. Pravin Malviya[2]

[1*2]*Departmen of Computer Science & Engineering, RGPV Bhopal India,*

Available online at: www.ijcseonline.org

Received:26/Jun/2016 Revised: 08/Jul/2016 Accepted: 12/Aug/2016 Published: 31/Aug/2016

Abstract— Now a day, large amounts of heterogeneous digital data is available this big data need to be carefully examined for analysis point of view. Big data is nothing but a large volume of heterogeneous and distributed data collection. In real world big data applications has contain huge amount of continuously grow able data but it is very costly to clean up, extract , manage and process data using present software tools. Fast and accurate retrieval of the relevant information from dataset has always been a significant issue. Prominent and accurate data clustering is a main task of exploratory data analysis and data mining applications. Clustering process is one of the data mining techniques for dividing informative dataset into group and it is a kind of unsupervised data mining technique.

Keywords— *Big data, Clustering*

I. INTRODUCTION

Today, big data is a main buzz word in domain of information technology, now new technologies of social communication driving the big data new trend as well as internet population grew day by day. Therefore, large companies like Facebook, Google, yahoo, YouTube etc. and for the purpose of analysis of this huge amount of data which is in unstructured form converted into prospered structured form. The big data analytics has following important parameters-variety, volume and velocity. [1]

Variety: The big data sources are extremely heterogeneous. The data files come in various formats and type, sometime it may be structured or unstructured such as text, log, audio, videos, files and many more.

Volume: Data volume is growing each day continuously by MB, KB, TB of information. The data results are stored into large files. Huge volume of data is an important issue of storage. This most important issue is resolved by reducing data storage cost. Data volumes are expected to grow 100 times within 10 years.

Velocity: The informative data comes at very high speed. Sometimes 60 second is too late so big data is very time sensitive. In some organizations data velocity rate is the main challenge.

Clustering Technique:

Data clustering is the most important task of data mining. It is an unsupervised method of machine learning application. In clustering the classes are divided according to class variable. Two important topics are:

1. Different ways to group a set of objects into a set cluster.

2. Types of clusters. The result of the cluster analysis is a number of heterogeneous groups with homogeneous contents. The first document or object of a cluster is defined as the initiator of that cluster. The initiator is called the cluster seed. Clustering algorithm design or selection: Patterns are grouped according to whether they resemble one another. The construction of a clustering criterion function makes the partition of clusters an optimization problem. [2] [3]

Clustering is used everywhere, and a wealth of clustering algorithms has been developed to solve different problems in specific domain. Therefore, it is important to carefully explore the characteristics of the problem on hand, in order to select or design an appropriate clustering strategy.

Cluster validation: Different approaches usually show the way to create different clusters and even for the same algorithm, types of parameter or the presentation order of the input patterns may influence the final result. Therefore, efficient evaluation standards and criteria are very important to provide the legitimate users with a degree of assurance, for the clustering results derived from the applied algorithms. Commonly, there are three categories of testing criteria: internal indices, external indices and relative indices. These are defined by following types of clustering structures, known as hierarchical clustering, partitioned clustering and individual clusters. [4] [5]

Result interpretation: The vital purpose of clustering is to provide users with meaningful insights into original data, so that they can effectively analyze and solve the problems encountered. Experts in the relevant fields can interpret the data partition. It may be required to guarantee the reliability and accuracy of the extracted knowledge.

Classification:

Classification is a simple but effective process to finding a model that describes and distinguishes data classes of test. Classification can be done by two types supervised learning and unsupervised. It consists of flowing steps:

Model Usage: This model is generally used unsupervised learning rule for defining future or unknown objects.

Model Construction: Consists of a set of predefined classes here the set of tuple used for model construction is known as training set. These type of models can be represented as classification rules, decision trees etc. [6]

II. PROBLEM DEFINITION

Big data clustering has been investigated for use in a number of different areas of text mining for prominent information retrieval. Initially, document clustering was investigated for improving the precision or recall in information retrieval systems. It is a very efficient way of finding the nearest neighbors of a document and applying proper clustering algorithms is also very important way of finding good clustering results. Unfortunately, there is little agreement over which is the best way to do clustering. The choice of evaluation methods frequently depends on the domain in which the research is being conducted. For example, an AI researcher might favor mutual information, while someone will choose F-measure. Two intuitive notions behind this is performance (accuracy) and precision which is need to be maintain while clustering any data set. [7][8]

III. PROPOSED METHODOLOGY

A. Big Data Technologies

Big data Test infrastructure requirement proper assessment. The various techniques and technologies have been introduced for cleaning, manipulating, analyzing and visualizing the big volume of data. There are many solutions to handle the big data but the Hadoop is one of the most widely used technologies currently used. But in this paper we are going to focus on only Map Reduce technique. [9]

B.Map Reduce

Map Reduce technology is a programming framework for distributed computing which is created by the Google Corporation, in which divide and conquer method is used to break the large complex data chunk and process them. Map Reduce: The master node takes the input. It divides into smaller subparts and distribute into worker nodes. A worker node further leads to the multi-level tree structure. The worker node processes the m=smaller problem and passes the answer back to the master node. The master node collects the answers from all the sub problems and combines together to form the output. [10][11]

C. Clustering Algorithm

K-means Algorithm:

This method is a type of hierarchical clustering method using K-means. The algorithm always starts by putting all the documents in a single cluster. It partitions the original clusters into multiple clusters by using K-means i.e. K=m. It makes the cluster which has highest intra cluster similarity as permanent & recursively split the other cluster into two more clusters using K-means with K=2 & continue this until and unless the desired number of clusters are created. [1]

Bisecting K-means Algorithm:

Bisecting k-means is most popular used algorithm to reduced dimensionality. Bisecting k-means is a combination of k-means as well as hierarchical k-means algorithm. It always starts with all objects in a single cluster. Bisecting K-means algorithm used for finding k-cluster.

Proposed Big-Data Mining Algorithm:

Step1: Input dataset for data cleaning and clustering process

Step2: Apply data cleaning process on selected dataset

Step3: Perform overall analysis on clean data set with various parameter

Step4: Pick any cluster to split.

Step5: Apply K-mean algorithm to find sub-clusters

Step6: Repeat Step4

The bisecting step for ITER times and takes the split that produces the clustering.

Step7: Repeat Step4,5, & 6 until the desired number of clusters are reached. [12][13]

IV. CONCLUSION

Big data processing framework needs to consider complex relationships between different samples, models and data sources. In big data mining high performance and accurate computing platforms are required. With big data analysis technologies, we will optimistically able to provide most accurate and most relevant social sensing feedback to our society at real-time. Map reduce mechanisms is suitable for large scale data mining process by testing series of standards data mining tasks on cluster. Map reduce implementation mechanism evaluated the proposed algorithm. A system needs to be appropriately designed so that unstructured data can be linked through their complex relationships to form useful informative patterns, the growth of data volumes, velocity and item relationships should help to from legitimate patterns.

REFERENCES

[1] BABU, G.P. and MARTY, M.N. 1994. Clustering with evolution strategies Pattern Recognition, 27, 2, 321-329.

[2] McKinsey Global Institute (2011) Big Data: The next frontier for innovation, competition and productivity.

[3] Shiv Pratap Singh Kushwah, Keshav Rawat, Pradeep Gupta" Analysis and Comparison of Efficient Techniques of Clustering Algorithms in Data Mining" International Journal of Innovative

[4] Chen, H., Chaing, R.H.L. and Storey, V.C. (2012) Business Intelligence and Analytics: From Big Data to Big Impact, MIS Quarterly, 36, 4, pp. 1165-1188.

[5] Neelamadhab Padhy, Dr. Pragnyaban Mishra and Rasmita Panigrahi, "The Survey of Data Mining Applications And Feature Scope", International Journal of Computer Science and Informatio Processing(CSIP).

[6] Wu Yuntian, Shaanxi University of Science and Technology, "Based on Machine Learning of Data Mining to Further Explore", 2012 International Conference on Machine Learning Banff, Canada.

[7] Guo, G, Neagu, D. (2005) Similarity-based Classifier Combination for Decision Making . Proc. Of IEEE International Conference on Systems, Man and Cybernetics, pp. 176-181

[8] Varun Kumar and Nisha Rathee, ITM University, "Knowledge discovery from database Using an integration of clustering and classification", International Journal of Advanced Computer Science and Applications, Vol. 2, No.3, March 2011.

[9] Wu, X., Zhu, X., Wu, G., Ding, W. (2014) Data Mining with Big Data, Knowledge and Data Enginnering , IEEE Transactions.

[10] Patel, A.B., Birla, M. and Nair, U. (2012) Addressing Big Data Problem Using Hadoop and Map Reduce, NIRMA University Conference on Engineering, pp. 1-5

[11] Aditya B. Patel, Manashvi Birla, Ushma Nair, (6-8 Dec. 2012),"Addressing Big Data Problem Using Hadoop and Map Reduce".

[12] Jyothi Bellary, Bhargavi Peyakunta, Sekhar Konetigari "Hybrid Machine Learning Approach In Data Mining", 2010 Second International Conference on Machine Learning and computing. Shiv Pratap Singh Kushwah, Keshav Rawat, Pradeep Gupta" Analysis and Comparison of Efficient Techniques of Clustering Algorithms in Data Mining" International Journal of Innovative.

[13] Fayyad, U. Data Mining and Knowledge Discovery: Making Sense Out of IEEE Expert, v. 11, no. 5, pp. 20-25, October 1996.