Various Data-Mining Techniques for Big Data

Manisha R. Thakare
Mtech (CSE)
B.D. College of Engineering,
Sevagram, Wardha, India

S. W. Mohod
Assistant Professor,
Department of C.E.
B.D. College of
EngineeringWardha,

A. N. Thakare
Assistant Professor,
Department of C.E
Engineering
B.D.C.E., Sevagram,
Wardha, India

ABSTRACT
Big data is the word used to describe structured and unstructured data. The term big data is originated from the web search companies who had to query loosely structured very large distributed data. Big Data is a new term used to identify the datasets that due to their large size and complexity. Big data mining is the capabilities of extracting useful information from these large datasets or streams data that due to its volume, variability and velocity. This data is going to be more diverse larger and faster.Mapreduce provides to the application programmer the abstraction of the map and reduce. Mapreduce is a framework used to write applications that process large amounts of data in parallel on clusters. Mapreduce framework for processing large amount of data. The main aim of this system is to improve performance through parallelization of various operations such as loading the data. This paper explores the efficient implementation of bisecting clustering algorithm with mapreduce in the context of grouping along with a new fully distributed architecture to implement the mapreduce programming model. The architecture also uses queries to shuffle results from map to reduce the cluster results also indicate that queues to overlap the map and shuffling stage seems to be a promising approach to improve mapreduce performance.

Keywords
Big data, Clustering, Classification, Clustering algorithms, Data Mining, Map-Reduce.

1. INTRODUCTION
1.1 Data Mining Techniques
Data mining having different type of techniques like clustering, classification, neural network etc but in this paper we are consider only two techniques i.e clustering and classification.[1] The information comes from heterogeneous, multiple, autonomous sources with their complex relationship. Big data growing up to 2.5 quintillion bytes of data are created daily and 90 percent data in the world today were produced within past two years. According to literature survey publicpicture sharing site it requires 1.8 million photos per day this shows that it is very difficult for big data applications to retrieve manage and process data from large volume of data. Currently big data processing depends upon parallel programming models like Mapreduce as well as providing computing platform of big data services. Data mining algorithms need to scan through the training data for obtaining the statistics for solving or optimizing model parameter.Data mining is an automated process used to extract valuable information from large and complex data sets. The several techniques in data mining classification and clustering are the main considerable point which is used to retrieve the essential knowledge from the very huge collection of data.

1.1.1 Clustering
Clustering is an unsupervised method of machine learning application. It is the most significant task of data mining. It is an unsupervised method of machine learning application.[19] Different ways to group a set of objects into a set cluster and 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. Feature extraction utilizes transformations to generate useful and novel features from the original ones. Feature selection chooses distinguishing features from a set of candidates. Ideal features should be of use in distinguishing patterns belonging to different clusters, immune to noise, easy to extract and interpret.[13] Clustering algorithm design or selection. The construction of a clustering criterion function makes the partition of clusters an optimization problem. Clustering is ubiquitous, and a wealth of clustering algorithms has been developed to solve different problems in specific fields. Cluster validation consists of different approaches usually lead to different clusters and parameter identification or the presentation order of the input patterns may affect the final result. Result interpretation contains the ultimate goal of clustering is to provide users with meaningful insights into original data, so that they can effectively solve the problems encountered. The relevant fields interpret the data partition. It may be required to guarantee the reliability of the extracted knowledge.[1]

1.1.2 Classification
Classification is a process to finding a model that describes and distinguishes data classes of test. It is types of supervised learning and unsupervised. The model construction consists of set of predefined classes. The set of tuple used for model construction is known as training set. This model can be represented as classification rules, decision trees. The model usage is used for defining future or unknown objects. It is used unsupervised learning rule. Classification data by using classification techniques in data mining is a very distinctive task. The first step is to build the model from the training set, i.e. randomly samples are selected from the data set. In the second step the data values are assigned to the model and verify the model’s accuracy.

2. PROPOSED METHODOLOGY
2.1 Big Data Technologies
The steps consist of big data test infrastructure assessment, infrastructure design, infrastructure implementation. The processing of large amount of data. The various techniques and technologies have been introduced for manipulating,
analyzing and visualizing the big data. There are many solutions to handle the Big Data but the Hadoop is one of the most widely used technologies. But in this paper we are consider only Map Reduce technique.

2.2 Map Reduce
Map Reduce is a programming framework for distributed computing which is created by the Google. The master node takes the input. It divides into smaller subparts and distribute into worker nodes. Mapreduce is a programming model which is inspired by functional programming and allows expressing distributed computations on massive amounts of data. It is an execution framework which is designed for large scale data processing run on clusters of commodity hardware.Mapreduce assists organizations in analyzing and processing for the multi-structured data of large volumes. Map Reduce has major application includes text analysis, machine learning, data transformation, indexing and search, graph analysis. Mapreduce has got a great popularity. It has the perception of data parallelism with a data model. A map-reduce framework put all pairs with the same key from all lists and gather them together. Mapreduce assume processing and storage nodes to be collocated. It is a feasible approach to tackle large data problems. It partitions a large problem into smaller sub-problems where independent sub-problems gets executed in parallel and combines intermediate results from each individual worker.

2.3 Clustering Algorithm:
2.3.1 Bisecting K-means Algorithm:
Bisecting K-Means is powerful, robust method which reduced dimensionality grouped into same cluster. Bisecting K-Mean gives better results for larger data sets than regular K-Mean. Bisecting K-Mean algorithm is less sensitizing to noise. Bisecting K-Means algorithm improvement in running time and accuracy highest mean clustering accuracy. It is relatively noise independent.

Steps in Bisecting K-Means Algorithm
1. Select K-points as the initial centroids.
2. Assign all points to the closest centroid.
3. Recomputed the centroid of each cluster.
4. Repeat 2, 3 until the centroids don’t change.

3. IMPLEMENTED MODULES
Dataset Used: The Airport dataset is used for performing training and testing of the system. The dataset from the KDD (Knowledge discovery and DM tools) Cup 1999. The dataset was downloaded and stored. It includes both training and testing records of airport datasets. The Airport dataset classified into different classes.

Data cleaning process i.e data cleansing/scrubbing frame. Data cleaning is the process of detecting and correcting (or removing) corrupt or inaccurate records from a record set, table or database. Extraction, transformation, loading include schema extraction and translation, schema matching and integration, schema implementation into instance extraction and transformation, instance matching and integration, filtering aggregation in data warehouse for scheduling, logging, monitoring, recovery, backup. Data warehouse is used for decision making i.e data vital to avoid wrong conclusion. Extract, transform, load refresh huge amounts of data into data repository for decision making. ETL software which including reading data from its source, cleaning it up, formatting it uniformly then writing it to the target repository.
needs to be moved from one place to another. The data lies in all sorts of heterogeneous system.

**Fig 5 : Different Cluster Analysis**
The big data analysis performs difference cluster analysis on the basis of overall analysis, country list, no. of time zone, no. of continent.

**Fig 6 : Different Analysis : Country List**
Country List contains country names with their etymologies including self governing political entity and tightly group of people which share a common culture.

**Fig 7 : Big Data Analysis : No.of Time Zone**
No.of time zone: A timezone is a region that has a uniform standard time for legal, commercial and social purpose (Local Time). Zones tend to follow the boundaries of countries and their subdivisions. Higher latitude countries use daylight saving time.

**Fig 8 : Big Data Analysis : No.of Continent**
No.of Continent: One of the seven great landmasses of the earth. Cluster Analysis depends on country list and timezone list. Search by no. of countries selection and no. of time zone and perform overall analysis. Screenshot shows Cluster Analysis depends on country list. Choose country or no. of countries and apply map reduce function. Mapping with google map i.e. connectivity with google map showing formation of cluster result. Perform Cluster analysis on the basis of country list.

**Fig 9 : Big Data Analysis Graph**
Analysis Graph (Over all Analysis): Graph shows value ranges from 0-8500 Records. The analysis graph describes on the basis of category and values are changes for number of records. Analysis graph indicates values ranges from 0 to 8500 Records. Create chart on the basis of domain axis label describes frequency data transfer object. The values provides on range axis label for frequency data transfer object.

**Fig 10 : Cluster Analysis : Country List**
Cluster Analysis depends on country list. Choose country or no. of countries and apply map reduce function. Mapping with google map i.e. connectivity with google map showing formation of cluster result. Perform Cluster analysis on the basis of country list. Cluster centroid countries.
3.1.1 Analysis Graph: Country List
The analysis graph shows cluster formation on the basis of search by two ways i.e. country list and timezone list. First search analysis graph on the basis of country list. The category indicates domain axis label and value ranges on axis label. The values denote the frequency data transfer object and route data transfer object.

3.1.2 Analysis Graph: Country List
The analysis graph shows cluster formation on the basis of search by two ways i.e. country list and timezone list. First search analysis graph on the basis of country list. The category indicates domain axis label and value ranges on axis label. The values denote the frequency data transfer object and route data transfer object.

4. CONCLUSION
The various data mining techniques for big data is computationally feasible for high dimensional datasets. The performance of clustering and classification is higher if it involves large dataset. The major strength of Data-Mining techniques is that the training of data is relatively easy. The feature selection is the process of selecting a specific subset of the terms of the training set and uses them in the classification. The feature selection process takes place before the training of the classifier. The main advantages for using feature selection algorithms are the facts that it reduces the dimension of our data, it makes the training faster and it can...
The proposed eering (ICDE), Systems, Man and l -ognition, Shaanxi University of Science and techniques in Data EE International Conference on, back to better understand our Mining IIJITEE) ISSN: 2278-54

...society at realtime. accurate social sensing feedback we will hopefully be able to provide most relevant and most computing platforms are required. With Big data technologies we will hopefully be able to provide most relevant and most accurate social sensing feedback to better understand our society at realtime.

5. REFERENCES