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 constructionconsists 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.Themodel usage is used for defining future or unknown objects. It is used unsupervised learning rule.Classifying 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. ROPOSED 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.[2] 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.[21]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.[20]

2.3 Clustering Algorithm:

2.3.1 Bisecting K-means Algorithm:

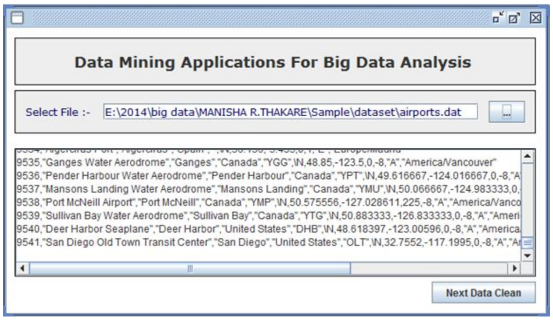
Bisecting K-Means is powerful, robust method which reduced dimentionality grouped intosame cluster. Bisecting K-Mean gives better results for larger data sets than regular K-Mean. Bisecting K-Mean algorithm is less sensitizing to noise. Bisecting K-Means algorithm improvement in running time and accuracy 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 centroidsdon'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.

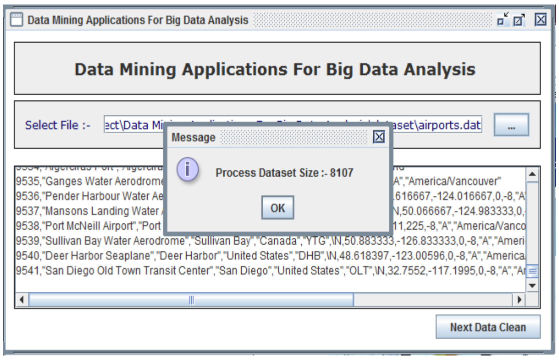


Data Mining Applications For Big Data Analysis

Select File :- E:\2014\big data\MANISHA R.THAKARE\Sample\dataset\airports.dat

```
9535,"Ganges Water Aerodrome","Ganges","Canada","YGG",\N,48.85,-123.5,0,-8,"A","America/Vancouver"
9536,"Pender Harbour Water Aerodrome","Pender Harbour","Canada","YPT",\N,49.616667,-124.016667,0,-8,"A
9537,"Mansons Landing Water Aerodrome","Mansons Landing","Canada","YMU",\N,50.066667,-124.983333,0,
9538,"Port McNeill Airport","Port McNeill","Canada","YMP",\N,50.575556,-127.028611,225,-8,"A","America/Vanco
9539,"Sullivan Bay Water Aerodrome","Sullivan Bay","Canada","YTG",\N,50.883333,-126.833333,0,-8,"A","Ameri
9540,"Deer Harbor Seaplane","Deer Harbor","United States","DHB",\N,48.618397,-123.00596,0,-8,"A","America
9541,"San Diego Old Town Transit Center","San Diego","United States","OLT",\N,32.7552,-117.1995,0,-8,"A","Ar
```

Fig 1 : Airport Dataset



Data Mining Applications For Big Data Analysis

Data Mining Applications For Big Data Analysis

Select File :- ect\Data Mi ... aset\airports.dat

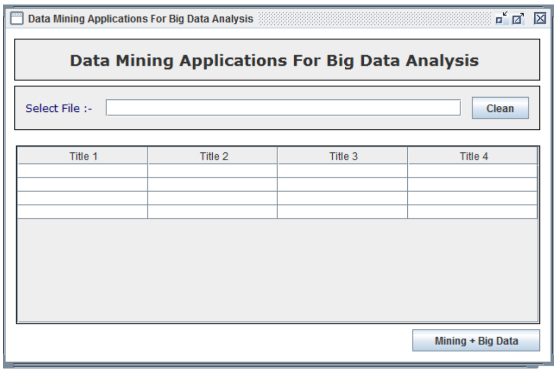
Message

Process Dataset Size :- 8107

```
9535,"Ganges Water Aerodrome                                    A","America/Vancouver"
9536,"Pender Harbour Water Ae                                    .616667,-124.016667,0,-8,"A
9537,"Mansons Landing Water                                     N,50.066667,-124.983333,0,
9538,"Port McNeill Airport","Port                               11,225,-8,"A","America/Vanco
9539,"Sullivan Bay Water Aerodrome","Sullivan Bay","Canada","YTG",\N,50.883333,-126.833333,0,-8,"A","Ameri
9540,"Deer Harbor Seaplane","Deer Harbor","United States","DHB",\N,48.618397,-123.00596,0,-8,"A","America
9541,"San Diego Old Town Transit Center","San Diego","United States","OLT",\N,32.7552,-117.1995,0,-8,"A","Ar
```

Fig 2 : Data Mining Applications for Big Data Analysis

The preprocess frame contains airport providing the airport ID, number, name along with the location, city, state. The dataset was downloaded and stored it includes both training and testing datasets. The process dataset size 8107 records of airplane flight aviation.



Data Mining Applications For Big Data Analysis

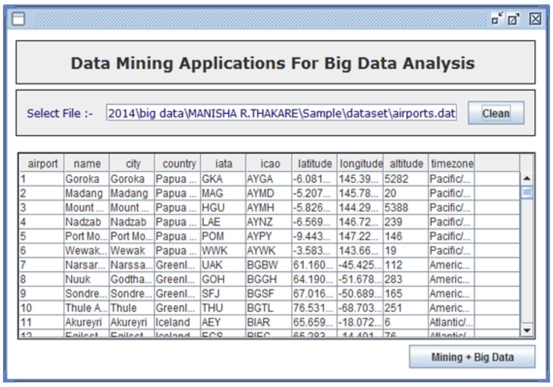
Data Mining Applications For Big Data Analysis

Select File :-

Title 1	Title 2	Title 3	Title 4

Fig 3 : Data Cleaning Frame

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.



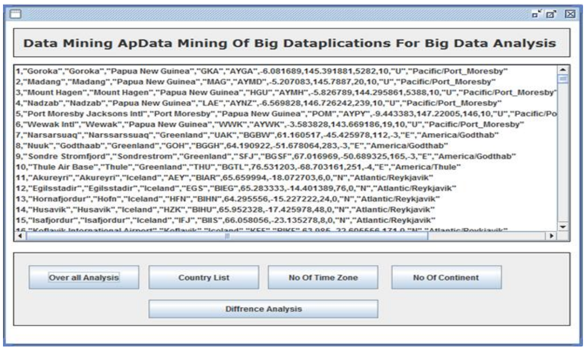
Data Mining Applications For Big Data Analysis

Select File :- 2014\big data\MANISHA R.THAKARE\Sample\dataset\airports.dat

airport	name	city	country	iata	icao	latitude	longitude	altitude	timezone
1	Goroka	Goroka	Papua ...	GKA	AYGA	-6.081...	145.39...	5282	Pacific/...
2	Madang	Madang	Papua ...	MAG	AYMD	-5.207...	145.78...	20	Pacific/...
3	Mount ...	Mount ...	Papua ...	HGU	AYMH	-5.826...	144.29...	5388	Pacific/...
4	Nadzab	Nadzab	Papua ...	LAE	AYNZ	-6.569...	146.72...	239	Pacific/...
5	Port Mo...	Port Mo...	Papua ...	POM	AYPY	-9.443...	147.22...	146	Pacific/...
6	Wewak...	Wewak	Papua ...	WWK	AYWK	-3.583...	143.66...	19	Pacific/...
7	Narsar...	Narssa...	Greenl...	UAK	BGBW	61.160...	-45.425...	112	Americ...
8	Nuuk	Godtha...	Greenl...	GOH	BGGH	64.190...	-51.678...	283	Americ...
9	Sondre...	Sondre...	Greenl...	SFJ	BGSF	67.016...	-50.689...	165	Americ...
10	Thule A...	Thule	Greenl...	THU	BGTL	76.531...	-68.703...	251	Americ...
11	Akureyri	Akureyri	Iceland	AEY	BIAR	65.659...	-18.072...	6	Atlantic/...

Fig 4 :Mining Big Data

The Airport dataset contains number of fields showing airport id, airport name, city, country, iata, icao, latitude, longitude, altitude, timezone. A process used by companies to turn raw data into useful information. Companies know they have valuable data lying around throughout their networks. That needs to be moved from one place to another. The data lies in all sorts of heterogeneous system.

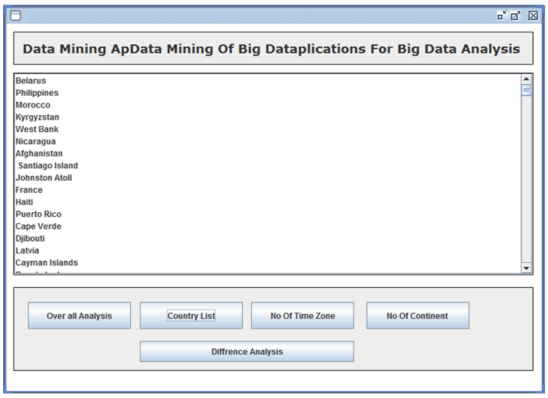


Data Mining ApData Mining Of Big Dataplications For Big Data Analysis

```
1,"Goroka","Goroka","Papua New Guinea","GKA","AYGA",-6.081689,145.391881,5282,10,"U","Pacific/Port_Moresby"
2,"Madang","Madang","Papua New Guinea","MAG","AYMD",-5.207083,145.7887,20,10,"U","Pacific/Port_Moresby"
3,"Mount Hagen","Mount Hagen","Papua New Guinea","HGU","AYMH",-5.826789,144.295861,5388,10,"U","Pacific/Port_Moresby"
4,"Nadzab","Nadzab","Papua New Guinea","LAE","AYNZ",-6.569828,146.726242,239,10,"U","Pacific/Port_Moresby"
5,"Port Moresby Jacksons Intl","Port Moresby","Papua New Guinea","POM","AYPY",-9.443383,147.22005,146,10,"U","Pacific/Po
6,"Wewak Intl","Wewak","Papua New Guinea","WWK","AYWK",-3.583828,143.669186,19,10,"U","Pacific/Port_Moresby"
7,"Narsarsuaq","Narssarssuaq","Greenland","UAK","BGBW",61.160517,-45.425978,112,-3,"E","America/Godthab"
8,"Nuuk","Godthaab","Greenland","GOH","BGGH",64.190922,-51.678064,283,-3,"E","America/Godthab"
9,"Sondre Stromfjord","Sondrestrom","Greenland","SFJ","BGSF",67.016969,-50.689325,165,-3,"E","America/Godthab"
10,"Thule Air Base","Thule","Greenland","THU","BGTL",76.531203,-68.703161,251,-4,"E","America/Thule"
11,"Akureyri","Akureyri","Iceland","AEY","BIAR",65.659994,-18.072703,6,0,"N","Atlantic/Reykjavik"
12,"Egilsstadir","Egilsstadir","Iceland","EGS","BIEG",65.283333,-14.401389,76,0,"N","Atlantic/Reykjavik"
13,"Hornafjordur","Hofn","Iceland","HFN","BIHN",64.295556,-15.227222,24,0,"N","Atlantic/Reykjavik"
14,"Husavik","Husavik","Iceland","HZK","BIHU",65.952328,-17.425978,48,0,"N","Atlantic/Reykjavik"
15,"Isafjordur","Isafjordur","Iceland","IFJ","BIIS",66.058056,-23.135278,8,0,"N","Atlantic/Reykjavik"
```

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.

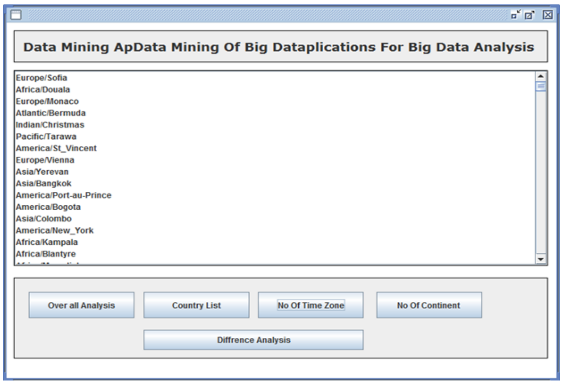


Data Mining ApData Mining Of Big Dataplications For Big Data Analysis

Belarus
Philippines
Morocco
Kyrgyzstan
West Bank
Nicaragua
Afghanistan
Santiago Island
Johnston Atoll
France
Haiti
Puerto Rico
Cape Verde
Djibouti
Latvia
Cayman Islands

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.

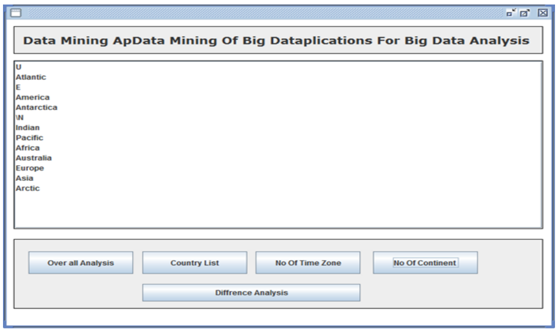


Data Mining ApData Mining Of Big Dataplications For Big Data Analysis

Europe/Sofia
Africa/Douala
Europe/Monaco
Atlantic/Bermuda
Indian/Christmas
Pacific/Tarawa
America/St_Vincent
Europe/Vienna
Asia/Yerevan
Asia/Bangkok
America/Port-au-Prince
America/Bogota
Asia/Colombo
America/New_York
Africa/Kampala
Africa/Blantyre

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.



Data Mining ApData Mining Of Big Dataplications For Big Data Analysis

U
Atlantic
E
America
Antarctica
\N
Indian
Pacific
Africa
Australia
Europe
Asia
Arctic

Fig 8 : Big Data Analysis : No.of Content

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.

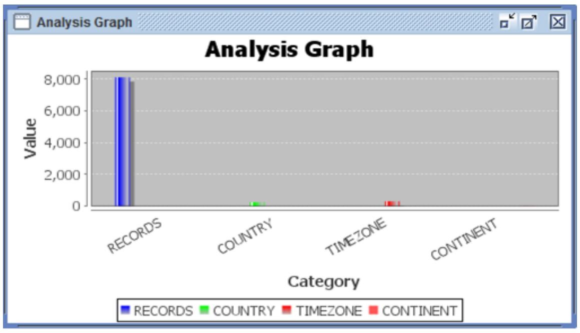
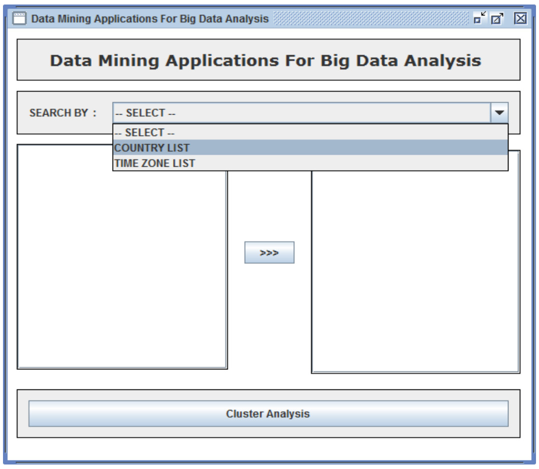


Fig 9 : Big Data Analysis Graph

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.



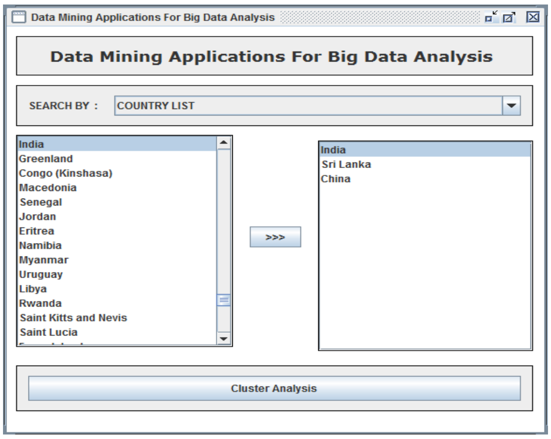
Data Mining Applications For Big Data Analysis

Data Mining Applications For Big Data Analysis

SEARCH BY : -- SELECT --

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



Data Mining Applications For Big Data Analysis

SEARCH BY : COUNTRY LIST

India
Greenland
Congo (Kinshasa)
Macedonia
Senegal
Jordan
Eritrea
Namibia
Myanmar
Uruguay
Libya
Rwanda
Saint Kitts and Nevis
Saint Lucia

India
Sri Lanka
China

Fig 11 : Cluster Analysis

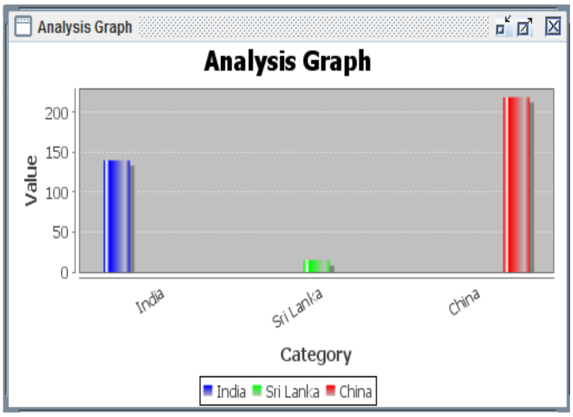


Fig 12 : Analysis Graph : Country List

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.

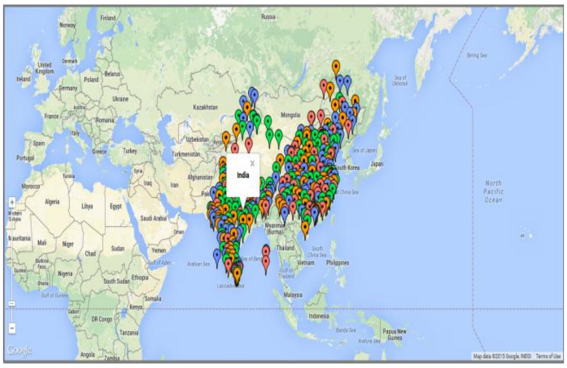
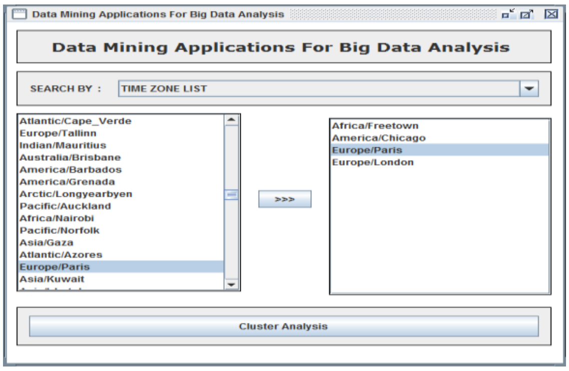


Fig 13 : Cluster Result



Data Mining Applications For Big Data Analysis

SEARCH BY : TIME ZONE LIST

Atlantic/Cape_Verde
Europe/Tallinn
Indian/Mauritius
Australia/Brisbane
America/Barbados
America/Grenada
Arctic/Longyearbyen
Pacific/Auckland
Africa/Nairobi
Pacific/Norfolk
Asia/Gaza
Atlantic/Azores
Europe/Paris
Asia/Kuwait

Africa/Freetown
America/Chicago
Europe/Paris
Europe/London

Fig 14 : Different Analysis : Time Zone List

Cluster Analysis depends on time zone 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 showing cluster centroid countries and their number of timezone.

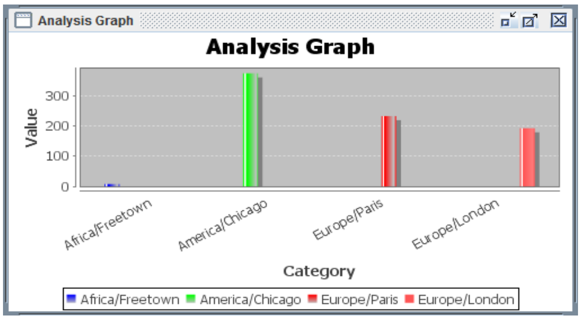


Fig 15 : Analysis Graph : No. of Time Zone

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.

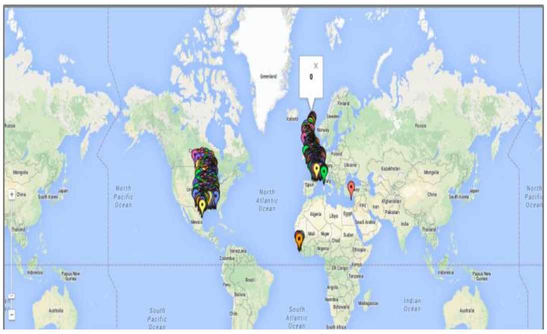


Fig 16 : Cluster Result

4. CONCLUSION

The various data mining techniques for big datais 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 improve accuracy by removing noisy features. The partition is done automatically by a clustering process.The proposed system achieved using technologies handle to big data i.e. MapReduce. The proposed system achieves different parameters of big data. Big data architecture contains several parts. Big data framework needs to consider complex relationships between samples, models and data sources along with their evolving changes with time and other possible factors. To support big data mining high performance 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

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

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

[3] 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

[4] Jyothi Bellary, BhargaviPeyakunta, SekharKonetigari "Hybrid Machine Learning Approach In Data Mining", 2010 Second International Conference on Machine Learning and computing.

[5] Oyelade, O. J, Oladipupo, O. O, Obagbuwa, I. C" Application of k- means Clustering algorithm for prediction of Students Academic Performance" (IJCSIS) International Journal of Computer Science and Information Security, Vol. 7, 2010.

[6] Varun Kumar and NishaRathee, 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.

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

[8] 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.

[9] 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.

[10] 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.

[11] NeelamadhabPadhy, Dr. Pragnyaban Mishra and RasmitaPanigrahi, "The Survey of Data Mining Applications And Feature Scope", International Journal of Computer Science and Information Processing(CSIP).

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

[13] Shiv Pratap Singh Kushwah, KeshavRawat, Pradeep Gupta" Analysis and Comparison of Efficient Techniques of Clustering Algorithms in Data Mining" International Journal of Innovative Technology and Exploring Engineering (IIJITEE) ISSN: 2278-3075, Volume-1, Issue-3, August 2012.

[14] Tekiner F. and Keane J.A., Systems, Man and Cybernetics(SMC), "Big Data Framework" 2013 IEEE International Conference on 13-16 Oct. 2013, 1494-1499.

[15] Dong, X.L.; Srivastava, D. Data Engineering (ICDE),' Big data integration" IEEE International Conference on, 29(2013)1245-1248.

[16] Sagiroglu, S.; Sinanc, D.,"Big Data: A Review",2013,20-24.

[17] Kyuseok Shim, MapReduce algorithms for Big Data Analysis, DNIS 2013, LNCS 7813, pp. 44-48, 2013.

[18] Madhuri V. Joseph, LipsaSadath and VanajaRajan" Data Mining: A Comparative Study on various Techniques and Methods" International Journal of Advanced Research in Computer Science and Software Engineering Volume 3, Issue 2, February 2013 ISSN: 2277.

[19] Aastha Joshi, RajneetKaur" A Review: Comparative Study of Various Clustering Techniques in Data Mining" International Journal of Advanced Research in computer Science and Software Engineering, Volume 3, Issue 3, March 2013.

[20] Yaxiong Zhao; Jie Wu INFOCOM, "Dache: A Data Aware Caching for Big-Data Applications Using the MapReduce Framework" 2014 Proceedings IEEE 2014, 35 – 39 (Volume 19).

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