International Journal of Scientific Research in Computer Science and Engineering
Vol.5, Issue.2, pp.9-13, April (2017)

Review Paper

E-ISSN: 2320-7639

Big Data Clustering Using Data Mining Technique

V.K. Gujare[1], P. Malviya[2]

[1]Department of Computer Science and Engineering, RGPV, Bhopal
[2]Department of Computer Science and Engineering, RGPV, Bhopal

**Corresponding Author: djvishal1234@gmail.com*

Available online at: www.isroset.org

Received 12 Feb 2017, Revised 25th Feb 2017, Accepted 18th Mar 2017, Online 30th Apr 2017

Abstract-- Big data is term is basically used for the collection of huge datasets, which may contain the structured or unstructured type of information. The data is growing day by day continuously. It is very difficult to manage and analyze such data and generate prominent and useful result from this type of data. Data mining is the technique used to analyze such data by applying different useful methods like preprocessing or cleaning, classification, clustering, prediction, transformation, selection, etc. So in this paper, we apply the data mining techniques on airport dataset to analyze the airport data and generate the result according to our need.

Keywords-- Big Data, Data Mining, Preprocessing, Classification, Clustering, 3v's of Big Data

I. Introduction

A. Big data

Big data is the term used for collection of huge amount of data. This term refers to data sets or combination of data sets whose size or volume, complexity /variability, and rate of growth /velocity make them difficult to be captured, managed, processed or analyzed by conventional technologies and tools, such as relational databases and desktop statistics or visualization packages, within the time necessary to make them useful.[1]

The two types of big data are : structured and unstructured.

Structured data are numbers and words that can be straightforwardly categorize and effectively analyzed. This type of are generated by things like, electronic devices, smart phones, and GPS etc. Structured data also provide information of account, sales figures and transaction type data.

Unstructured data include more complex information, like review of customer from some commercial websites, images and other multimedia information, collected comments on famous social networking sites. These data cannot easily be separated into categories or analyzed numerically.

"Unstructured big data is the things that humans are saying," says big data consulting firm vice president Tony Jewitt of Plano, Texas[2, 3]. It uses natural languages communication data like tweeter or face book etc Analysis of such unstructured data relies on important keywords, which allow users to clean the data based on desirable searchable terms.

B. Big data Characteristics:

Big data starts with large-volume, heterogeneous, autonomous sources with distributed and decentralized control, and seeks to explore complex and evolving relationship among data.[2]Big data can be described by the following characteristics.

Volume

It is amount of data generated and stored. The size/volume of the data determines the value as well as potential insight- and whether it can actually be considered as a big data or not.

Variety

Describe the type and nature of the data. This helps the researcher to analyze it to effectively use the resulting insight.

Velocity

In this context, the data speed at which it is generated as well as processed to meet the required demands and challenges that present in the path of growth and development.

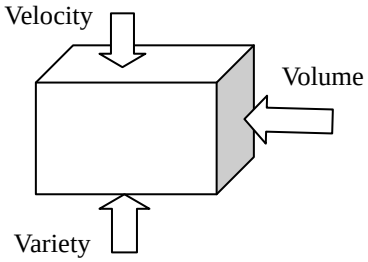

Fig1: Depicts three V's in Big Data

In this paper, the airport dataset is used to show the working of data mining techniques on airport dataset. To manage the

unstructured data, we use the preprocessing technique like stop word removal.

II. DATA MINING

Data mining process refers to extracting knowledge from relevant large amounts of data. Mostly people used data mining term as an alternative for another popularly used term, i.e. Knowledge Discovery from data. On the other hand, data mining is simply an essential and effective step in the process of knowledge discovery:

A. Data Cleaning :-To Remove noise and inconsistent data.
B. Data integration :- Where multiple data sources may be combined.
C. Data selection: - Here data which is relevant to the analysis task are extracted from the database to perform analysis on that data.
D. Data transformation :- Where data are transformed or consolidated into from appropriate for mining by mining by performing summary or aggregation operation for instance.

III. CLUSTERING

Clustering can be treated as the most important unsupervised learning method. Thus every other problem of this kind, it deals with seeking a structure in a collection of unlabeled data.

Clustering is a process of grouping particular type of objects based on their observable characteristics and aggregating them according to their similarities. Regarding to data mining, this methodology partitions the data clustering implementing a specific join algorithm, most suitable for the desired information analysis.

This clustering analysis allows an object not to be part of a cluster, or strictly belong to it, calling this type of grouping hard partitioning [6].

A. Related Work
In the project, the data set of airport system is used to perform data mining techniques and analyze the data to get the result of number of airports present in a particular country with location. To process data, different processes are used such as preprocessing, data mining, classification, clustering. After performing data mining techniques, we will get the result.

B. Big data
Big data consists of structured and unstructured type of data. Structured data contains the numbers and words that can be easily categorized and analyzed.Unstructured datainclude more complex information, such as customer reviews from commercial websites, photos and other multimedia, and comments on social networking sites.
Big Data starts with large-volume, heterogeneous, autonomous sources with distributed and decentralized control, and seeks to explore complex and evolving relationships among data[1].These characteristics make it an extreme challenge for discovering useful knowledge from the Big Data.[4]

Analyzing big data is a very tedious task. Analyze the big data, there are different techniques that can be used like preprocessing, data mining, classification, clustering, etc.
In this project, we are taking the airport dataset as a example of big data. On that data, we will apply the different techniques to get the result which is required by user. To preprocess the data, we are implementing stop word removal technique.

After pre-processing, we will perform the classification technique by implementing k-NN algorithm.
After classification, we will perform the clustering technique by implementing k-means algorithm.
Now, we will display the result in the form of graphs and the result will display the exact location of the particular airport in map with the help of Google map.

C. Data Cleaning
Mostly real world data or information is noisy, incomplete and inconsistent. Data cleaning process is used to fill missing values, smooth out real time noise while identifying unwanted outliers, and correct inconsistencies in the given data.

Stopword removal technique
For cleaning process, there are different algorithms can be used. Data cleaning is the important process of data mining. Therefore, to clean the data in this project, we are using the stopword removal technique. The data is present in the form of inconsistent, irrelevant format. To remove these inconsistency, in these project we are using the stopword removal technique.

In stopword removal technique, the irrelevant data is removed from the dataset. The airport dataset contains the inconsistent data with commas (,), double- quotation (“ “), etc. Therefore, after applying this technique, the irrelevant data is removed from the dataset. Again, the dataset contains, the irrelevant fields, i.e., some values are not present in the fields. These fields are also removed from the dataset.

Working of stopword removal technique:
Step 1: Reading of dataset record by record.
Step 2: If the records contains fields less than 8 parameters then that type of data is discarded from dataset.

Step 3: Again, if the record contains commas (,), double-quotation (“ “) then also the record get removed from dataset.
Step 4: It will save the data in a tabular format.

D. Data Classification Technique:

K-NN algorithm

KNN is a *non parametric lazy learning* algorithm. That is a pretty concise statement. When you say a technique is non parametric, it means that it does not make any assumptions on the underlying data distribution. This is pretty useful, as in the real world, most of the practical data does not obey the typical theoretical assumptions made (eggaussian mixtures, linearly separable etc). Non parametric algorithms like KNN come to the rescue here.

It is also a lazy algorithm. This means the training phase is pretty fast .Lack of generalization means that KNN keeps all the training data. More exactly, all the training data is needed during the testing phase. (Well this is an exaggeration, but not far from truth). This is in contrast to other techniques like SVM where you can discard all non support vectors without any problem.

The steps of the K-nearest neighbors are as follows:

Step 1:-Determine the parameter value of K where K is number of nearest neighbors beforehand. This value is depending on user.

Step 2:- Calculate the distance between the data-instance and all the training samples here we can use any distance algorithm.

Step 3:- Sort the distances for all the samples and find the nearest neighbor based on the K^{th} minimum distance.

Step 4:-This process uses supervised learning, so get all the categories of your training sample data for the sorted value which fall under K. Use the majority of nearest neighbors as the prediction value.
In the project, when we apply the classification technique, the input data will be classified into different groups according to country list, time-zone, and components.

E. Data Clustering Technique:

This clustering analysis allows an object not to be part of a cluster, or strictly belong to it, calling this type of grouping hard partitioning. In the other hand, soft partitioning states that every object belongs to a cluster in a determined degree [6]. More specific divisions can be possible to create objects that belonging to multiple clusters, to force an object to participate in only one cluster or even construct hierarchical trees on group relationships.

K-means algorithm:

The algorithm proceeds by finding the distance between each data point is assigned to a cluster belonging to the closet centroid. In the next step the centroids are recomputed by taking the mean value of all the data points in a cluster. This process is repeated till the centroids no longer move more than a specified threshold value.

K – means clustering algorithm:
Step 1:-Start with k centroid points
Step2:-While the centroids no longer move beyond a threshold or maximum number of iterations reached;
Step 3:-for each point in the dataset;
Step 4:- for each centroid:
Step 5:-find the distance between the point and the centroid
Step 6:-recompute the centroid point by taking mean value of all points in the cluster.

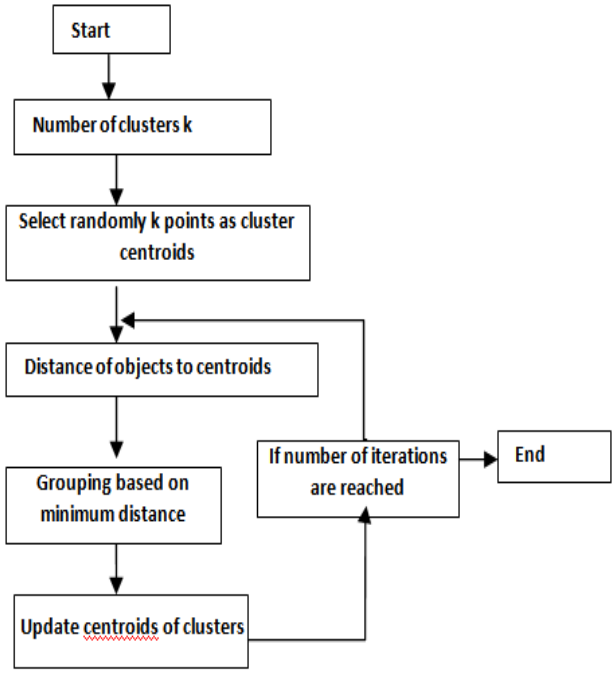

Fig 2 :- Working of K – means clustering algorithm

Implemented work and Experimental Analysis
In the project, for the cleaning technique, the stopword removal technique is used.In stopword removal technique, the irrelevant data is removed from the dataset.

To implement the Classification here , k-NN algorithm is used .K-nearest neighbor algorithm (KNN) is part of supervised learning that has been used in many applications in the field of data mining, statistical pattern recognition and many others[6].

For Clustering, k-means algorithm is used.The *k*-means method is not guaranteed to converge to the global optimum and often terminates at a local optimum.
To display the result of analysis, we are using the graphs pattern. To display graph, we are using JFreeChart package

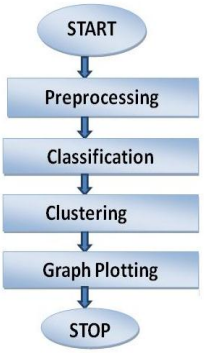

Fig 3: Implemented Work Flow

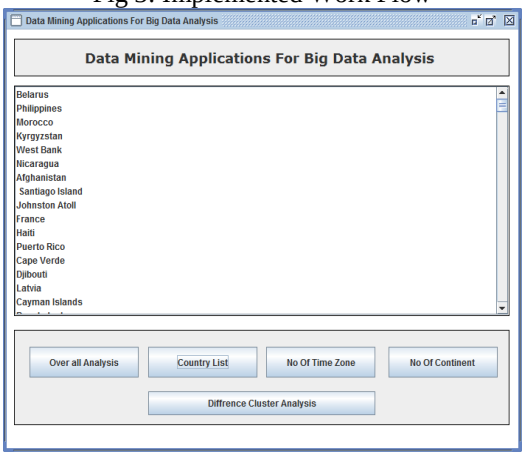

Fig 4:- Classification of Data

This is the result of classification technique. Here, we are classifying the data according to no of countries. We can classify the data according to the no of time zones and no of continents.

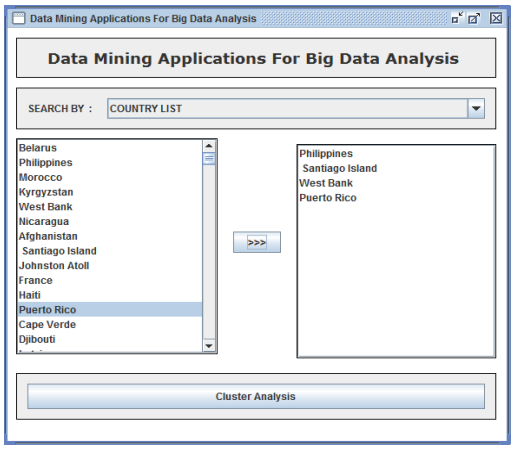

Fig 5 :- Cluster Analysis According To Country List

Here, we are performing clustering technique by using k-means algorithm. We will make the cluster according to country names or time zones. It will provides the number of airports present in the selected countries or time zones.
In the following examples, we are making the clusters according to country list using bisecting k-means algorithm. Here, we have selected four countries.

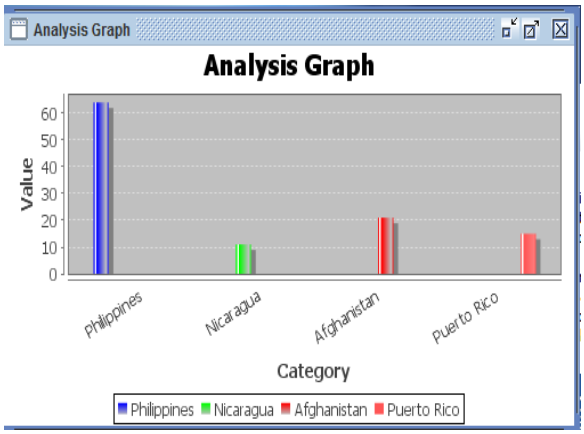

Fig 6 :- Cluster Analysis Graph According To Country List

This is the output of clustering technique. Here, the number of airport present in Countries are represented in the graph format.

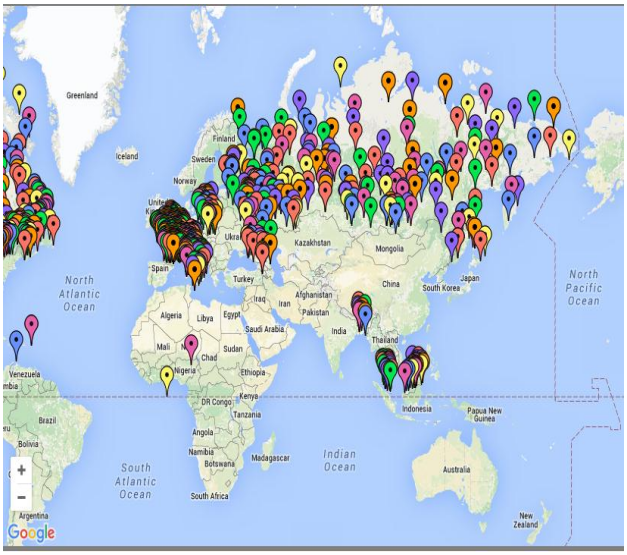

Fig 7 Location of airports using google map

This page is shows the final result. Here, the location of airports of selected countries are represented using google map.

IV. Conclusion

Today, the Big Data term is used for the large amount of data sets. The dataset a collection of huge number of records this huge and complex records are becomes difficult to process using old tool or traditional data processing application. Therefore, processing the data becomes tedious task. Extracting knowledge from large data, we are performing data mining technique in this process we discovering interesting knowledge from large amount of data stored in airport dataset, data warehouse or other information. To analyze the working of data mining process on big data, we are using the airport dataset as an example of big data. We are applying different techniques to process the data such as preprocessing, data mining, classification, clustering, etc. We are making the clusters of airports according to country and time-zone. We get the number of airports present in country in a graphical format. It shows the specific location of airports of particular country using google – map.

References

[1]. P. Sharma, V. Garg, R. Kaur, S. Sonare, "*Big Data in Cloud Environment*", International Journal of Computer Sciences and Engineering, Vol.1, Issue.3, pp.15-17, 2013.

[2]. X. Wu , G.Q. Wu, W. Ding, "*Data Mining with Big data*", IEEE Transactions on Knoweledge and Data Engineering, Vol 26, No.1, 2014

[3]. M. Dhivya, D. Ragupathi, V.R. Kumar, *"Hadoop Mapreduce Outline in Big Figures Analytics"*, International Journal of Computer Sciences and Engineering, Vol.2, Issue.9, pp.100-104, 2014.

[4]. M.S. Begamr, N. Vetrivelanr, "*An Analysis on Data Mining Processes on Big Data Framework*", International Journal of Computer Sciences and Engineering, Vol.2, Issue.12, pp.76-79, Dec -2014

[5]. M. Kumarasamy, G. N. K. Suresh Babu, "*Applications of Big Data in various Domains*", International Journal of Computer Sciences and Engineering, Vol.4, Issue.5, pp.81-85, 2016.

[6]. C. Blake and C. Merz, "*UCI repository of machine learning databases*", University of California, California, pp.1-16, 1998.

