IJARCCE

ISSN (Online) 2278-1021
ISSN (Print) 2319 5940

International Journal of Advanced Research in Computer and Communication Engineering
Vol. 5, Issue 1, January 2016

A Paramount Mining for Facebook Dataset using Hadoop.

Vrushali R Choudhari[1], Suchita Walke[2]

M.E Student, Department of Computer Engineering., SESOI, Bhivpuri, India[1]

Head of Department, Information Technology, YTCM, Bhivpuri, India[2]

Abstract: This Big Data is new term define to store a huge data sets complex data. There are various techniques which are proposed for mining. The big data can be mined is really a researchers issue especially in domain of text mining. This paper presents an effective technique which include a preprocessing of huge dataset(i.e. text mining) for finding the shortest neighbor. the dataset used here is facebook .The proposed system provides a solution for storing huge data and retrieving which is adapted to all environment.

Keywords: component; Dataset, Preprocessing, Clustering, Big Data, K-Means.

I. INTRODUCTION

Big Data application are the collection of data which are large and are beyond the ability of software tools. Data Complexity of industries are exponentially growing and storing of such enormous data has become archaic. The Nostrum for all companies is an hadoop software.Hadoop software has become innate nowdays.The big data provides solutions and build large scale processing system. Data Mining involves expanding and analyzing huge data to find various patterns of big data. The source for big data can be social networking data, publicily available sources and streaming data.

In Social network data we can say it as social media data. The data which is on social interaction is increasingly attractive set of information particularly for sales, marketing and support .So it can be unstructured or semi-structured form.

It poses a unique challenges when it comes to consumption and analysis. After identifying all potential source of data we shall start harnessing of information as how to store and manage it,how much of it analyze etc. In business many companies can collect terabytes or petrabytes of information using big data.

Data Mining is an optimistic and passably new technology. Data Mining is defined as a process of recognize hidden patterns and knowledge from database using different techniques such as machine learning, artificial intelligence and statistical. Data Mining is a vital part of Knowledge discovery process we can evaluate an enormous set of data and get unrevealed and useful knowledge. It include generalization, classification, clustering, association, pattern matching, data visualization, characterization are examined. Data Mining is also used in transportation, insurance, government, weather forecast, medicine etc. Data Mining allows a search for precious information in database. The military use data mining to learn different play in the accuracy of bombs. Intelligence agencies might use it to determine which large quantity of intercepted communications .Medical researchers use for predicting cancer relapse.

Different techniques for searching data and building models are there in statistics: Linear regression, logistic regression, discriminant analysis and principal component analysis.

Data mining can be used in Marketing /Retail which bring lot of assets to retail in the same way of industries. Data mining also provide information related to loan and credit reporting for financial institutions. They are applying data mining in operational engineering data. In government sectors for analyzing records to detect criminal activities or money laundering data mining is used. The key challenges and research issues include: - designing Big Data sampling mechanisms to reduce Big Data volumes to a manageable size for processing; - building prediction models from Big Data streams. Such models can adaptively adjust to the dynamic changing of the data. A knowledge indexing framework to ensure real-time data monitoring and classification for Big Data applications.

II. PROPOSED SYSTEM

First, In this we collect dataset of facebook from the UCI machine learning repository website.After collecting the dataset for this process.Then we convert the dataset into csv file format.In this proposed system after preprocessing and clustering of data.We get outlier of processed data then mean value is calculated.In proposed system we get the shortest neighbor of facebook user.

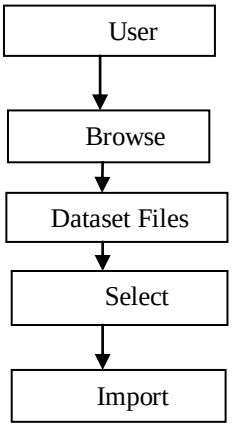
A. Dataset :

The Dataset is collection of information. We can say that it contains single database table where each column of table represents a variable and each row has given dataset. Every Dataset is having value which is called as “datum”.

In this we collect a dataset from UCI machine learning.UCI (university of California) Irvine which has collection of data. Each domain or field is preferred in dataset which has list of data to be processed out. In the proposed system we have taken the facebook dataset from UCI .The downloaded dataset has comma separated value file format.The.CSV saves the data in structured data

IJARCCE
ISSN (Online) 2278-1021
ISSN (Print) 2319 5940
International Journal of Advanced Research in Computer and Communication Engineering
Vol. 5, Issue 1, January 2016

format. It stores the data in tabular form in plain text or in database. Each line in file is called as data record.



B. PreProcessing

Preprocessing includes removing garbage values, removing contaminated data and special symbols. The Preprocessed data is extracted and processed for big detection. We can use string tokenization for preprocessing .

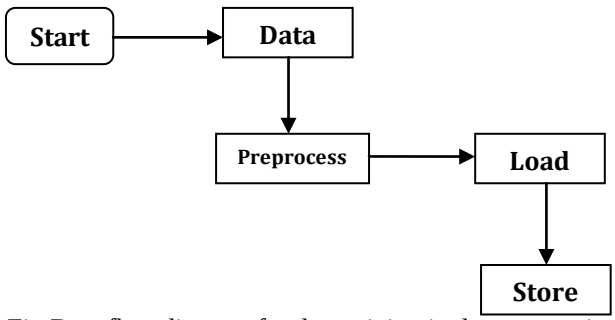


Fig.Data flow diagram for data mining in data processing technique based on data set.

C. . Clustering :

In proposed system we are calculated the overlap value for each and every attribute present in dataset. Clustering algorithm and cluster validity are analogize parts in cluster analysis. A centroids Ratio is used for comparing different clustering results.The exchange strategy in the algorithm for simple perturbation to solution and concurrent for nearest vale by K-means.

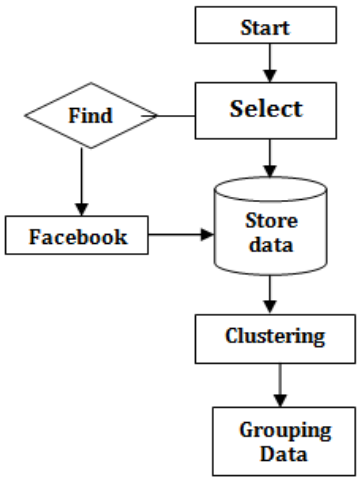


Fig. Clustering

III .K- MEANS

It creates 'K' groups from set of objects. so that members of group are more similar. K-means is technique required for exploring a dataset. Cluster analysis are designed to form groups such that group members are more similar versus non-group member.

For example : Suppose we have dataset of facebook. In cluster analysisi these would be called as observations. we can define different things as Name, age, Sex, date of birth, sex, college, education etc. This is vector representing the facebook.

K-means has lots of variations to optimize for certain types of data. Vector as list of numbers. We know about the facebook. List can also be interpreted as coordinates in multi-dimensional space. Name can be one dimensional ,dateofbirth can be another dimension and so forth. Given set of vectors how we do cluster together facebook that have similar name, college, age etc.

K-means has lots of variations to optimize for certain type of data.

K-Means Algorithm :

i. KK-means selects points in multi-dimensional space to represent each of k- cluster. They are called as centroids.
ii. EE very User will be nearest to one of the centroids. As users won't be nearest to near one, so they shall form a cluster to their closet centroids.
iii. EE very User is now member of a cluster so now we have K- Clusters formed.
iv. KK-Clusters members (as such Name, Dateofbirth, age, Education, college etc) are used to perceive the center using this algorithm.
v. T The Forge center becomes new centroid for cluster.
vi. C Cluster membership can be changed as it is located at different place now user can be nearby to other centroids.
vii. R Repeat steps ii to step vi until centroids no longer revamp and stiffen cluster membership. This is called as Concurrence.

IV. CONCLUSION

In Real World application and Organizations ,Agencies ,managing and mining big data is provocative but a enthralling tasks. Big Data analyze data volumes. Data mining techniques are widely for Business Intelligence. It is the science of customer relationship management. It is concluded that development of data mining techniques is tending to extract information from data.

Data mining brings a lot of benefits to businesses, society, governments as well as individual. However privacy, security and misuse of information are the big problems if they are not addressed and resolved properly.To Evaluate Big Data we have perused various data ,model and system levels.

Nowdays Big Data as transpire and growing in all science and engineering domains. To understand at real time big data provide pertinent and integrate social sensing feedback.

DOI 10.17148/IJARCCE.2016.5159

REFERENCES

[1] Xindong Wu,Fellow,"Data Mining With Big Data "
[2] R.Chen ,k.Sivakumar"Collective Mining of Bayesian Networks from Distributed Hetrerogeneous Data"
[3] S. Aral and D. Walker, "Identifying Influential and Susceptible Members of Social Networks," Science, vol. 337, pp. 337-341, 2012.
[4] A. Machanavajjhala and J.P. Reiter, "Big Privacy: Protecting Confidentiality in Big Data," ACM Crossroads, vol pp. 20-23, 2012.
[5] Dataminingruhul.blogspot.in/2011/03/architecture-of-typical-data-mining html.
[6] Adderley, R., Townsley, M., & Bond, J. (2007). Use of data mining techniques to model crime scene investigator performance. Knowledge-Based Systems, 20(2),170–176

DOI 10.17148/IJARCCE.2016.5159