www.ijecs.in
International Journal Of Engineering And Computer Science ISSN:2319-7242
Volume 6 Issue 10 October 2017, Page No. 22797-22801
Index Copernicus value (2015): 58.10 DOI: 10.18535/ijecs/v6i10.15

Data Mining in Bigdata using RCDC Clustering Algorithm

B NagaLakshmi[1], Nagi Setty A[2]

NIIT Pvt Limited,Hyderabad, NeuDesic Pvt Ltd Hyderabad.

Abstract:
Big Data refers to the large set of data that range beyond exa-bytes(1018).It is due to the exceeding capacity of online storage and processing system. The evolution of web technology results in a huge amount of data present in the web for the internet user. Those data's exist in the form of text, audio, video, image, etc. Text Mining, knowledge discovery in text (KDT) is the process of extracting knowledge from the huge amount of data that has been available in the form of text. The rapid growth of internet usage in the social media leads to the generation of more volume of data. It has been estimated that 80% of the business information are available in the form of unstructured and semi-structured text data. K-means is the most common clustering algorithm used to cluster the similar content in hadoop environment. Since it is based on "bag of words" the lexical semantic analysis is not used. Our proposed RCDC algorithm is based on latent semantic analysis, which is more efficient, scalable and accurate.

***Key Words*:** HDFS, MapReduce, K-means, RCDC, WordNet Ontology.

1. Introduction

The term Big Data evolved due to the unstructured and semi-structured data that are available in various formats. Those data cannot be processed in Relational database management.nMany open source tools such as Hadoop, Hive, Pig, Sqoop, Cassandra, etc are available to process those big data. Here, we are focusing on hadoop environment to process the data as it is robust and allows distributed
processing. i.e, the big data applications will continue to run even when individual servers or clusters fail. In hadoop, the input data are stored in hadoop distributed file system (HDFS), which is based on the GFS (Google File System). These data are processed in
the MapReduce concept that maps and reduces the data as needed. Basically K-means is the clustering algorithm used in the MapReduce concept which is inefficient. Relative concept document clustering (RCDC) algorithm makes use of WordNet Ontology lexical database used to find the related words and cluster it.

The medical data are scattered in both structured and unstructured format .The data available in medical industry includes genomes, proteomics, metabolomics and common diagnosis of patients. The advantage here is prevention and intervention of diseases. [1] The issue involved in it is privacy and security.

The term big data involves the innovative technologies to store, manage, retrieve, and distribute a large set of data. Big data is composed of structured, un-structured and semi-structured data, which is difficult for processing.[2] The conventional data management system is inefficient to process these un-structured data, so Hadoop acts as a core platform for big data and perform the processing of data.

The clustering algorithm is the grouping of similar data into a cluster. This paper performs an analysis of all the clustering algorithms such as K-Means, Bisecting K-Means and Machine Learning in data mining.[3] It describes the advantages and shortcomings of each algorithm. The issues involve time consumption, complexity and accuracy.

The text document contains many useful patterns which have been discovered and mined by many data mining techniques. This paper focuses to effective use and update of useful discovered patterns by text

mining techniques.[4] The advantages over here are pattern deploying and evolving and refine the discovered patterns. The issue identified here is low frequency problem.

2. **System Implementation**

K-means is the most common clustering algorithm used to cluster the similar content in the text documents, which it is based on "**bag of words**".

The existing system uses K-Means in java and MapRedK-Means to cluster the large set of input documents in space.

K-Means clustering algorithm generates a specific set of disjoint and non-hierarchical clusters. It is well suited for generating globular clusters. The K-Means uses numerical, unsupervised, non-deterministic and iterative method. K-Means performs vectorization, by means of converting the

clustering, but it is not suited for non-globular data and consumes more time.

The traditional K-Means clustering algorithm is tedious for non-globular data and the cluster centers are created randomly. This makes the clusters susceptible to noise points and is very unstable. MapReduce K-Means is based on the parallel processing of data by means of K-Means algorithm and MapReduce concept, mapper() and reducer() functions. The vectorization of all the documents in space is achieved by means of vector class in K-Means algorithm**.** The ClusterCenter finds the centroid value for all the generated vectors. The Distance between the vector and centroid is calculated based on the Euclidean distance using DistanceMeasurer. Mapper class available in the MapReduce, performs the series iterations to create the clusters based on the distance between the centroid and vector. Reducer creates the cluster until the convergence is achieved. The Job Configuration is used to execute the MapReduce job of K-Means.

The traditional K-Means and MapReduce K-Means are inefficient for clustering of documents in big data, as it consumes more time and does not uses lexical semantic analysis. Our proposed system RCDC, relative concept based document clustering uses WordNet Ontology to find the relationship between the data. It is used to place all the closely related information under same clusters and somewhat related information under nearby clusters and not-related information under far away clusters.

3. **Hadoop**

text document into numerical values known as vectors. A cluster center known as centroid is randomly created for the set of vectors. The distance between the centroid and the vectors is calculated and the tightly packed clusters are created. It is suited for globular clusters than hierarchical

It is one of the open source platform developed by Apache software foundation in order to analyze the bigdata which is fully written in java. It is useful to store huge amount of data efficiently and cheaply.[13] Many vendors are available for the processing of Bigdata since; Hadoop is the grandfather of all. I.e. it is used to store enormous amount of data in distributed clusters and it is designed as robust even if one individual server or clusters fail. Hadoop mainly consists of two parts such as distributed file system called HDFS and data processing framework called MapReduce.

HDFS is the Hadoop Distributed File System based on Google File system which is used to store huge set of data sets to process [9]. Even though it is not a database, the input files that to be processed are stored in HDFS.

Those files are either in the format of either .txt or .csv files. The input files are divided into default file size of 64MB. It provides replication of files, such that the files are stored in different cluster or servers. Usually it will store the files in three individual servers.

HDFS consists of five nodes to store the input files [9]. In that the first three nodes acts as Master Node and others as Slave node. They are:

Name Node: like a meta data used to indicate the files located in data node

Secondary Name Node: supporting name node Job Tracker: distributes individual tasks to machines running the task tracker

Data Node: used to store the input files in different nodes of 64MB block size by means of replicationMapreduce:

Task Tracker: monitor the jobs of individual machines.

MapReduce is found by Google employees Jeffrey Dean and Sanjay Gheemawat which is used in Google to process the file system available to the users. MapReduce is based on java code such that it is used to process the files in HDFS.

data files and reduce () is used to combine the outputs from taking the results of a mapper. The input to the mapper is processed based on the key-value pair [9]. The data types of key value pair has to be indicated for both the input and output of the file. Since, the input and output variable data types may vary. The input to the MapReduce is the files stored in HDFS such that these are divided into n-splits. The job tracker in the HDFS is responsible for creating the splits from the files stored in HDFS. The Record Reader is used to transform the split into key-value pairs.Used by Map Reduce

DOI: 10.18535/ijecs/v6i10.15

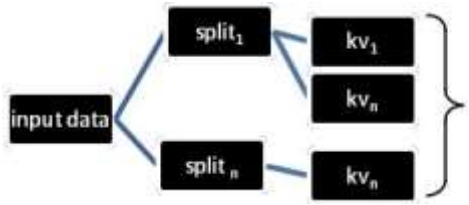

Fig -1 : Mapreduce Input

4. System Architecture

The input is taken from the different source and it may vary from stream to stream and from user to user. If the input is from medical oriented then the users are doctors. If the input is from educational institution then the users will be students and members of the educational institution. These inputs are in the form of unstructured data in which it can be processed by the big data platform Hadoop.

These inputs are stored in Hadoop HDFS file system in which it is used to store huge amount of file. The files in HDFS are divided into small number of blocks each of size 128 MB or 64 MB. These files are stored in replicated in three nodes to provide this as a redundancy and available.

These files are given as an input to MapReduce which is used to map and reduce the input files. The basic K-means algorithm is followed here in which the centroid value is used to find the clustering algorithms. The map generates a key-value pair as an output which is taken as input to the reduce phase and the final key value pair is taken as an output.

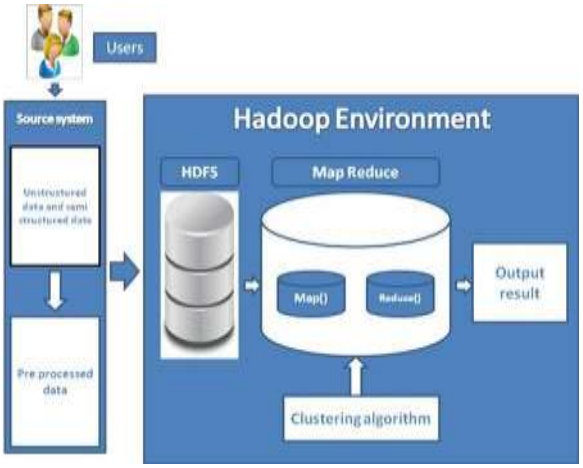

Fig -2 : System Architecture

5. Analysis of Algorithm

K-means is the basic clustering algorithm which is based on Lloyd algorithm which identifies the cluster center based on Euclidean distance and it is iterative. This algorithm is widely used because of its simplicity and it is used to process huge amount of data sets. Here the centroid is generated for each cluster and the distance between the centroid and the nodes are calculated. The number of files given as an input is taken as k. For each k vector is represented and the vector is taken as a centroid value. The clustering is done by measuring the similarity or distance between the documents. The Euclidean distance is the basis for finding out the difference between the documents.

The distance between the documents is measure by the following formula,

$$J=\sum_{i=1}^{k}\sum_{j=1}^{n}\|xi-cj\|_2$$

Where, Xi is the number of clusters

Cj is the centroid of clusters

The k-means algorithm is implemented as:

A. Calculate k randomly selected points from n points as centroid.

B. Randomly assign a centroid point to each cluster based on the Euclidean distance.

C. Compute the distance for all clusters based on the centroid.

D. Redo steps 2 and 3 until there will be no change in points.

The bisecting K-means algorithm is one of the clustering algorithms used in hadoop MapReduce architecture in which it is based on both the K-means and the hierarchical clustering algorithm. The main advantage over the bisecting K-means is it is used to reduce the additional disk space needed. I.e. it is used to minimize the wastage of space.

Basic Bisecting K-means algorithm for finding K clusters

a. Pick a cluster to split
b. Find 2 sub cluster based on Bisecting K-means
c. Repeat step 2 for iteration times and take the split that produces clustering with the overall similar.
d. Repeat the steps 1,2,3 until the expected clusters reached.

The critical part is determining which cluster is choose for splitting [6]. And there are different ways to proceed, for example, you can choose the

biggest cluster or the cluster with the worst quality or a combination of both.

The Mapreduce K-means is based on the K-Means algorithm done by mapping and reducing concept. The vectors are the spatial representation of the documents. That is the input documents are processed and stored as a vector value. The centroid value has to be defined for for the vectors[12]. The distance between the centroid and the vector is achieved by Euclidean distance and the clusters are created. Hence the mapreduce K-means algorithm is based on the K-means algorithm. The Mapper is used to find the cluster center for the vectors created in the space. The cluster centroid and the distance for each iteration is monitored and viewed by the mapper that is created. The Reducer is used for maintaining the number of iterations to be performed. The iterations are done until the convergence is achieved.

Our proposed RCDC algorithm is Relative Concept Document Clustering which is based on lexical semantic analysis such that the relationship between the words is identified. The lexical semantic analysis is used to identify the relationship between the words used that are tokenized. Instead of using bag of words for clustering RCDC algorithm looks for the related words which gives the latent semantical meaning for the word. The input is taken from the HDFS file system for processing such that RCDC gets the key value pair that to be found out. Then the tokenizations of words are processed and the documents in the space are identified in the basis of lexical semantic analysis .

The words that are related together can be grouped together as a bag of words are clustered. The documents are first identified based on the lexical semantic analysis and it is represented as points in space. Then the cluster center vectors for each vectors in space has to be calculated. The Euclidean distance between the vectors and the clustered center has to be measured and the closest point has to be identified. If not the convergence value is met then it has to be iterated for multiple times.

6. Experimental Results

K-Means is the most common clustering algorithm in which the input data is processed as “bag of words” and the clusters are created, which consumes more time and not suited for non-globular data. In Mapred K-Means, the input data is processed simultaneously by K-Means and MapReduce algorithms. The time consumption is less when compared with K-Means clustering. Our proposed system, RCDC uses relative concepts and lexical semantic analysis instead of using bag of words. Different words with similar meaning is converted into same word with same meaning and counted as one. Thus the clusters are created in much easier way. This reduces the time consumption and creates tight clusters when compared with all other algorithms. It is efficient, scalable and accurate.

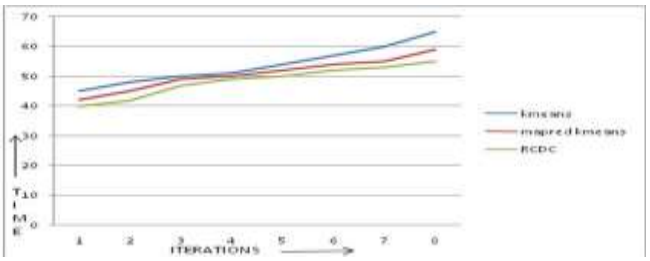

Fig -3 : Performance analysis of RCDC vs Mapred KMeans vs KMeans

7. Conclusion

With the increasing size of the data worldwide from various enterprise applications, it has become quite critical to perform data analysis over the massively growing data. Hence, text-mining in big data offers greater deal of flexibility for an organization to extract latent information from the massively growing textual data and thereby evolve up with much better informative values. Although Bigdata has recently gained more attention from last 5 years, but still there are many issues in this research domain. This paper has reviewed some of the clustering algorithm used. The unstructured text is very common, and it represents the majority of information available. The text mining in Bigdata has providing the typical applications such as analyzing open-ended surveys, automatic processing of messages and emails, analyzing warranty or insurance claims and investigating the competitors by crawling in their web sites.

REFERENCES

1. Javier Andreu-Perez, Carmen C. Y. Poon, Robert D. Merrifield, Stephen T. C. Wong, and Guang-Zhong Yang “Bigdata For health” IEEE Journal, 2015.

2. Harshawardhan S. Bhosale1, Prof. Devendra P. Gadekar2 "A review paper on Bigdata and hadoop" ISRP Journal, Volume 4, Issue 10, October 2014.

3. Mythili S, Madhiya E "An analysis clustering algorithms in Data Mining" Journal IJCSMC, Vol. 3, Issue. 1, January 2014, pg.334 – 340.

4. Ning Zhong, Yuefeng Li, and Sheng-Tang Wu "Effective Pattern discovery for Text Mining" IEEE Journal, VOL. 24, NO. 1, JANUARY 2012.