Extended K-mean clustering technique using Map Reduced Segmentation clustering for Big Data Analytics

M.V.S. Prasad1,, Dr. O.Naga Raju2
1Research Scholar, Department of Computer Science and Engineering,
Acharya Nagarjuna University Andhra Pradesh, India -522510
2Lecturer in Computer Science, Ph.D & HoD of Computer Science,
S. K. B. R Govt. Degree College, Macherla, Guntur, Andhra Pradesh, India -522426

Abstract : Now a days internet of things becoming one among the foremost important sources for data as these data could also be utilized in tons of application inside smart city which can help to form the lifetime of the human less difficult and cozy. The demand of knowledge mining methods to realize tons of data from this valuable source becomes more vital. data processing algorithms should be processed via using suitable computing technique like distributed computing. Distributed computing may be a model wont to do high computational processing over a group of connected systems. Each individual system interconnected on the network is named a node and therefore the collection of the many nodes that form a network is named a cluster. Clustering is an important method of knowledge analysis through which the first data set are often partitioned into several data subsets consistent with similarities of knowledge points. It becomes an underlying tool for outlier detection, biology, indexing, and so on. within the context of fuzzy clustering analysis, each object in data set not belongs to one group but possibly belongs to any group.

Apache Hadoop is an open-source software framework that supports data-intensive distributed applications, licensed under the Apache v2 license. It supports the running of applications on large clusters of commodity hardware. Hadoop was derived from Google's MapReduce and Google filing system (GFS)

I. INTRODUCTION

With the rapid development of mobile Internet, cloud computing, Internet of things, social network service, and other emerging services, data is growing at an explosive rate recently. the way to achieve fast and effective analyses of knowledge then maximize the info property's benefits has become the main target of attention. The "four Vs" model, variety, volume, velocity, and value, for giant data has made traditional methods of knowledge analysis inapplicable. Therefore, new techniques for giant data analysis like distributed or parallelized, feature extraction, and sampling are widely concerned.

The Hadoop framework transparently provides both reliability and data motion to applications. Hadoop implements a computational paradigm named MapReduce, where the appliance is split into many small fragments of labor, each of which can be executed or re-executed on any node within the cluster. additionally, it provides a distributed filing system that stores data on the compute nodes, providing very high aggregate bandwidth across the cluster. Both map/reduce and therefore the distributed filing system are designed in order that node failures are automatically handled by the framework.

1.2. Cluster Analysis

Data mining is interdisciplinary topic which might be defined in many different ways. There are variety ofdata mining methods are accustomed determine the kinds of patterns to be found in data processing task. These methods include discrimination and characterizations, frequent patterns mining, correlations and associations, classification and regression; clustering analysis, outlier analysis.

Clustering is one among the foremost exciting topics in data processing. Clustering utilized in many application areas like business intelligence, image pattern recognition, biology, security, and Web search. the target of clustering is to explore intrinsic structures in data, and arrange them into expressive subgroups. the fundamental concept of cluster analysis is that the process of dividing large data set of objects into small subsets. Each small subset may be a single cluster, such the objects are clustered together looking on the concept of minimizing interclass and maximizing the intraclass similarity. Similarity and dissimilarity are assessed supported the feature values describing objects and various distance measures. We measures object's similarity and dissimilarity by comparing objects with one another. These measures include distance measures like supremum distances, Manhattan distance, and Euclidean distance, between two objects of numeric data. Cluster analysis could be a vast topic and hence there are many clustering algorithms available to group datasets[1].

On the basis of implementation different clustering algorithm can be grouped together into

Partitioning Method
- K-means
- K- medoids Hierarchical Method
- Chameleon

- BIRCH

Density Based Clustering Method

- OPTICS
- DBSCAN

Grid Based Clustering Method

- CLIQUE
- STING

1.3 Hierarchical Clustering algorithms:

There are two approaches to perform Hierarchical clustering techniques Agglomerative (top-bottom) and Divisive (bottom-top). In Agglomerative approach, initially one object is chosen and successively merges the neighbor objects supported the space as minimum, maximum and average. the method is continuous until a desired cluster is made. The Divisive approach deals with set of objects as single cluster and divides the cluster into further clusters until desired no of clusters are formed. BIRCH, CURE, ROCK, Chameleon, Echidna, Wards, SNN, GRIDCLUST, CACTUS are a number of Hierarchical clustering algorithms during which clusters of Non convex, Arbitrary Hyper rectangular are formed[2].

Density based Clustering algorithms:

Data objects are categorized into core points, border points and noise points. All the core points are connected together supported the densities to create cluster. Arbitrary shaped clusters are formed by various clustering algorithms like DBSCAN, OPTICS, DBCLASD, GDBSCAN, DENCLU and SUBCLU.

Grid based Clustering algorithms:

Grid based algorithm partitions the information set into no number of cells to create a grid structure. Clusters are formed supported the grid structure. to create clusters Grid algorithm uses subspace and hierarchical clustering techniques. STING, CLIQUE, Wave cluster, BANG, OptiGrid, MAFIA, ENCLUS, PROCLUS, ORCLUS, FC and STIRR. Compare to any or all Clustering algorithms Grid algorithms are in no time processing algorithms. Uniform grid algorithms don't seem to be sufficient to make desired clusters. to beat these problem Adaptive grid algorithms like MAFIA and AMR Arbitrary shaped clusters are formed by the grid cells[3].

1.4 .Methodology

Big Data Analytics

Big data analytics is that the process of examining big data to find hidden patterns, unknown correlations and other useful information that may be accustomed make better decisions. To perform any reasonably analysis on such large and complex data, scaling up the hardware platforms become necessary and selectingthe correct platforms becomes a vital decision to satisfy the user's requirement in fewer amounts of your time. There are various big data platforms available with different characteristics. to decide on a right platform for specific application one should have knowledge of the benefits and limitations of of these platforms. The platform you select must be ready to cater to increased processing demands if it's appropriate to create the analytics based solutions on a selected platform[5]. This data comes from many various sources: The smart phones, the info they generate and consume; sensors embedded into everyday objects, which resulted in billions of latest and constantly updating data feed containing location, climate and other information; posts to social media sites, digital photos and videos and get transaction records. This data is termed big data. the primary organizations to grab it were online and startup firms. Firms like Facebook, Google and LinkedIn are built around big data from the start. "Big Data" refers to data sets overlargeand sophisticated containing structured, semi-structured and unstructured data, which is extremely difficult to handle with traditional software tools. In many organizations, the quantityof knowledgeis greater or it moves faster or it exceeds current processing capacity. An example of massive data could be Petabytes (1,024 terabytes) or Exabyte's (1,024 petabytes) of information containing billions to trillions of records of many various users—all from different sources like social media, banking, web, mobile, employees and customer's data etc.Thesevarieties of data are typically loosely structured data that's often incomplete and inaccessible.

K-means algorithm

The Lloyd's algorithm, mostly called k-means algorithm, is employedto resolve the k-means clustering problem and works as follows. First, decide the amount of clusters k. Then: Clustering is that the process of partitioning a gaggleof information points into alittle number of clusters. as an example, the thingsduring a supermarket are clustered in categories (butter, cheese and milk are grouped in dairy products). in factthis can be a qualitative quite partitioning. A quantitative approach would be to live certain features of the products, say percentage of milk et al., and products with high percentage of milk would be grouped together. generally, we've n data points xi,i=1...n that need to be partitioned in k clusters[6]. The goal is to assign a cluster to everydatum. K-means may be a clustering method that aims to seek out the positions μi,i=1...k of the clusters that minimize the space from the info points to the cluster. K-means clustering solves

$$\operatorname{argminc} \sum i=1k \sum x \in cid(x, \mu i)=\operatorname{argminc} \sum i=1k$$

$$\sum x \in ci // x-\mu i // 22$$

whereci is the set of points that belong to cluster i. The K-means clustering uses the square of the Euclidean distance $d(x,\mu i)= // x-\mu i // 22$. This problem is not trivial(in fact it is NP-hard), so the K-means algorithm only hopes to find the global minimum, possibly getting stuck in a different solution.

The K-means clustering algorithm is process by usingMapReduce can be divided into the following phases:

1. Initial

- The given input data set can be split into sub datasets. The sub datasets are formed into <Key, Value> lists. And these <Key, Value> lists input into map function.
- Select k points randomly from the datasets as initial clustering centroids.

2. Mapper

a) Update the cluster centroids. Calculate the distance between the each point in given datasets and k centroids.
b) Arrange each data to the nearest cluster until all the data have been processed.
c) Output <ai, zj> pair. And ai is the center of the cluster zj.

3. Reducer

- Read <ai, zj> from Map stage. Collect all the data records. And then output of k clusters and the data points.

- Calculate the average of each cluster which is selected as the new cluster center.

Initializing the position of the clusters
It is really up to you! Here are some common methods:

1. Forgy: set the positions of thekclusters to k observations chosen randomly from the dataset.
2. Random partition: assign a clusterrandomly to each observation and compute means as in step

3. Since the algorithm stops in a local minimum, the initial position of the clusters is very important.
The pseudo code for k-means clustering algorithm is given below:
Input: Data points D, numbers of clusters k Step 1: Slaves read their part of data
Step 2: do until global centroids to the slaves Step 3: Master broadcasts the centroids to the slaves
Step 4: Slaves assign data instances to the closest centroids
Step 5: Slaves compute the new local centroids and local cluster sizes
Step 6: Slaves send local centroids and cluster sizes to the master
Step 7: Master aggregates local centroids weighted by local cluster sizes into global centroids.
Output: Data points with cluster memberships.

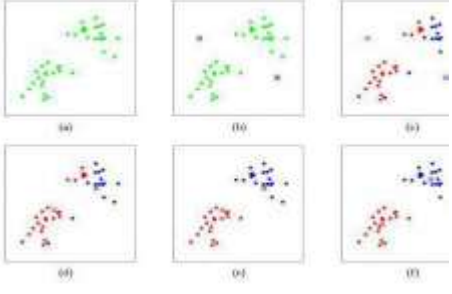

Figure 1.1: K-means algorithm.

Training examples are shown as dots, and cluster centroids are shown as crosses. (a) Original dataset. (b) Random initial cluster centroids. (c-f) Illustration of running two iterations of k-means. In each iteration, we assign each training example to the closest cluster centroid (shown by "painting" the training examples identical color because the cluster centroid to which is assigned); then we move each cluster centroid to the mean of the points assigned thereto. In the clustering problem, we are given a training set x(1),...,x(m), and wish to group the information into a couple of cohesive "clusters." Here, we are given feature vectors for everydatumx(i)∈R n as usual; but no labels y(i) (making this an unsupervised learning problem). Our goal is to predict k centroids and a labelc(i) for everydatapoint.

1.5 . AN OPTIMIZED K-MEANS CLUSTERING USING MAP-REDUCE TECHNIQUE

The first step of designing MapReduce code Kmeans algorithm is to specific and investigate the input and output of the implementation. Input is given as pair, where “key” is that the cluster mean and “value‟ is that theserializable implementation of a vector within the dataset. The prerequisite to implement Map routine and Reduce routine is to possess two files. the primary one should involve clusters with their centroids values and therefore the other one should have objects to be clustered. Chosen of centroids and therefore the objects to be clustered are arranged in two spillited files is that the initial step to cluster data by K-means algorithm using MapReduce method of Apache Hadoop [7]. It may be done by following the algorithm to implement MapReduceroutines for K-means clustering. The initial set of centroid is stored within the input directory of HDFS before Map routine call and that they form the “key‟ field within the pair. The instructions required to computethe space between the given data set and cluster centroid fed as a pair is coded within the Mapper routine. The Mapper function calculates the gap between the thing value and every of the cluster centroidreferredwithin the cluster set and jointly keeping track of the cluster to which the given object is closest.Once the computation of distances is complete the thing should be assigned to the closest cluster. Once Mapper is invoked, the given object is assigned to the cluster that it's nearest associated with. After the assignment of all objects to their

associated clusters is completed the centroid of every cluster is recomputed.The recalculation is completed by the Reduce routine and also it restructures the cluster to avoid generation of clusters with extreme sizes. At the top, once the centroid of the given cluster is revised, the new set of objects and clusters is re-written to the memory and is prepared for subsequent iteration.

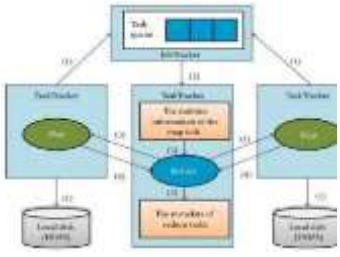

Figure 1.2: Acquisition of the metadata for reduce tasks.

1.6 . Repartitioning

The repartitioning process divides the collected virtual partitions into new partitions of identical number as reduce tasks. the info size of the largest partition are often minimized after repartitioning process. It also can reduce the interval needed for the utmost partition, thereby speeding up the completion of the whole reduce phase and increasing the speed of completed jobs also as system throughput. As previously analysed, the repartitioning process recombines each virtual partition generated within the map phase. However, thanks to the limitation of accessible memory, these virtual partitions must be written to the local classification system. If repartitioning isn't restricted, it's likely to steer to a plurality of discrete virtual partitions in one partition following the balancing process, leading to a non-sequential read of the disk.

Algorithm 1: Repartitioning algorithm.
Data: A a1,a2,…,an,K
Result: R: an index of subsequenceStep1:low←max ai
Step 2:high←n
Step 3:num←0
Step 4:whilelow<highdo
Step 5:mid←low + high − low /2 Step 6:foreachai∈Ado
Step 7:sum←sum + a1 Step 8:ifsum>midthen Step 9:num+ +
Step 10:sum←a1
Step 11:R←R∪i Step 12:end Step 13:end
Step 14:ifnum ⩽ Kthen Step 15:high←mid − 1 Step 16:end
Step 17:elseifnum>Kthen
Step 18:low←mid + 1 Step 19:end
Step 20:end
Step 21: return R;

1.7. Comparison of Clustering Algorithms
Volume:
It refers to the power of an algorithm to affect large amounts of a knowledge. With reference tothe quantity property the standards for clustering algorithms to be considered may be a. Size of the info set b. High dimensionality. Handling Outliers. Size of the info set: Data set is collection of attributes. The attributes are categorical, nominal, ordinal, interval and ratio. Many clustering algorithms support numerical and categorical data.

High dimensionality: To handle big data as the size of data set increases no of dimensions are also increases. It is the curse of dimensionality.Outliers: Many clustering algorithms are capable of handle outliers. Noise data cannot be making a group with data points.

Variety:

Variety refers to the ability of a clustering algorithm to handle different types of data sets such as numerical, categorical, nominal and ordinal. A criterion for clustering algorithms is (a) type of data set (b) cluster shape.

Type of data set: The size of the data set is small or big but many of the clustering algorithms support large data sets for big data mining.

Cluster shape: Depends on the data set size and type shape of the cluster formed.

Velocity:

Velocity refers to the computations of clustering algorithm based on the criteria (a) running time complexity of a clustering algorithm.

Time complexity: If the computations of algorithms take very less no then algorithm has less run time. The algorithms the run time calculation done based on Big O notation.

Value:

For a clustering algorithm to process the data accurately and to form a cluster with less computation input parameter are play key role.

1.8 .MapReduce Processing Model

HadoopMapReduce processes big data in parallel and provides output with efficient performance. Map-reduce contains Map function and Reduce function. Map function executes filtering and sorting of huge data sets. Reduce function performs the summary operation which mixes the result and provides the improved output. Hadoop HDFS and Map-Reduce are delineated with the assistance of Google filing system. Google filing system (GFS) is developed by Google may be a distributed filing system that provide organized and adequate access to data using large clusters of commodity servers. Map phase: The Master node accepts the input then divides an outsized problem is into smaller sub-problems. It then distributes these sub-problems among worker nodes during a multi-level tree structure. These sub-problems are then processed by the worker nodes which execute and sent the result back to the master node. Reduce phase: Reduce function combines the output of all sub problems and collect it in master node and produces final output. Each map function is related to a reduce function.

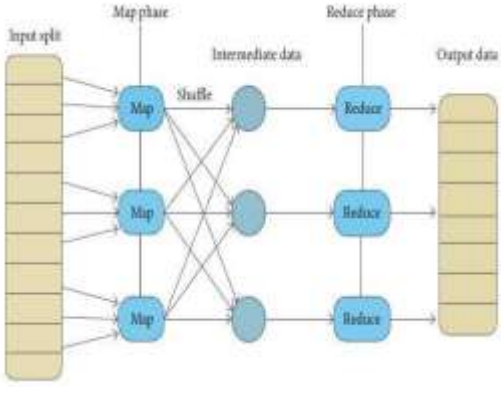

Figure 1.3: Map Reduce Programming Model

Operation mechanism of MapReduce is as follows:

(1)Input: MapReduce framework based on Hadoop requires a pair ofMap and Reduce functionsimplementing the appropriate interface or abstract class,and should also be specified the input and output location and other operating parameters.In this stage, the large dataintheinput directory will be divided into several independent data blocks for the Map function of parallel processing.

(2)MapReduce framework puts the application of the input as a set of key-value pairs <key, value>. In the Map stage, the framework will call the user-defined Map function to process each key-value pairs<key, value>, while generating a new batch ofmiddlekeyvalue pairs<key, value>.

(3)Shuffle：In order to ensure that the input of Reduceoutputted by Map have been sorted, in the Shuffle stage, the framework uses HTTP to get associated key-value pairs <key,value> Map outputs for each Reduce; MapReduceframeworkgroups the input of the Reduce phase according to the key value.

(4)Reduce：This phase will traverse the intermediate data for each unique key, and execute user-defined Reduce function. The input parameter is < key, {a list of values} >, the output is the new key-value pairs< key, value >. (5)Output：This stage will write the results of the Reduce to thespecified output directory location.

1.9 . Results and Discussion

Experimental setup

To implement the k-means algorithm we installed cluster composed of four nodes in aws,

3. One Master Node (instance) of typem4.
4. Four slave Nodes (instance) of type m4.

3 Hadoop 2.4.1

4. JDK 1.7

This distributed environment of four instances in AWS used to implement, perform the optimized repartitioned k-means clustering algorithm and to save the results.

Data set description

To scale optimized repartitioning k-means clustering algorithm one of smart city dataset used . We used the pollution data set which consists of 449 file. Each file contains around 17500 observation of the pollutants ratio of five attributes.

Evaluation

To measure the performance of the scaled k-means algorithms using HadoopMapReduce, we have executed thealgorithms on 10 different samples of data. After execution of the algorithm, we have calculated and measure the inter-cluster and intra-cluster similarity measure.

The inter-cluster distance: distanced(i,j)between two clusters is measuredas the distance between the centroids of the clusters.

The intra-cluster distance measuredbetween the all pair of objects within a cluster.

The following table and figure represent the experimental results of K-means algorithm on different data samples where k=3.

Table 1.1: Execution results of An optimized repartitioned K-means cluster algorithm

Sample	Sample size	Inter-Cluster Density	Intra-Cluster Density
S1	78290	0.689142	0.556309
S2	1576718	0.740337	0.561887
S3	2368512	0.73014	0.562767
S4	3153530	0.748691	0.5684
S5	3942470	0.802399	0.567079
S6	4732842	0.676724	0.611366
S7	5522887	0.74722	0.563842
S8	6312932	0.783907	0.565958
S9	7099392	0.704998	0.56797
S10	7887974	0.771926	0.572288

Figure 1.4 : Comparison between Inter-cluster and Intra Cluster

From the results it is clear the sample s5 shows the maximum inter-cluster density of 0.552399 which indicates well separation of different cluster. Similarly, the inter-cluster density for sample s8 is calculated as 0.533907, separating data clusters very well. Also the results of Intra-cluster density for sample s1 show minimum value, which gives a clear indication of having the similar objects in the same cluster.

1.10.Conclusion

In this thesis we propose an Optimized repartitioned k-means clustering algorithm scaled up to be applied to large dataset which contain around 10 millionobjects. Each object may be a vector of six attributes. Inter and intra cluster measurements computed to seek outthe utmost value of inter-cluster density and therefore the minimum value of intra-cluster measurements. This research work done using Hadoop and MapReduce framework which provides high performance in big data analysis.

References

[1] Murty, M.N. Clustering large data sets, Soft Computing approach to Pattern Recognition and Image Processing 53, 41-63 (World Scientific, Singapore, 2002) ISBN: 981-238-251-8

[2] Jain, A. K. and Dubes, R.C. Algorithms for clustering data (Prentice Hall, New Jersey, 1988)

[3] Fasulo, D. An Analysis of Recent Work on Clustering Algorithms (1999), http://citeseer.ist.psu.edu/fasulo99analysi.html

[4] MacQueen, J. Some methods for classification and analysis of multivariate observations, Proc. 5 th Berkeley Symp. Math. Stat. and Prob., 1, 281-97 (Univ. of Calif. Press, 1967)

[5] Anderberg, M.R. Cluster Analysis for Applications (Academic Press, New York, 1973)

[6] Everitt, B.S. Cluster Analysis (John Wiley & Sons, New York., 1974)

[7] Kaufman, L. and Rousseeuw, P.J. Finding Groups in Data. An Introduction to Cluster

Analysis (John Wiley & Sons, Canada, 1990)

[8] Berkhin, P. Survey of Clustering Data Mining Techniques (2002), http://citeseer.ist.psu.edu/berkhin02survey.html

[9] Jain, A. K. Data Clustering: 50 Years beyond K-Means, Pattern Recognition Letters. 31, 8, 651-666. (2009)

[10] Su, T. and Dy, J. G. In search of deterministic methods for initializing K-means and Gaussian mixture clustering, Intelligent Data Analysis 11 (4), 319-338, (IOS Press, 2007)

[11] Bradley, P. S. and Fayyad, U. M. Refining Initial Points for K-Means Clustering, Proc. 15th Intl Conf on Machine Learning, 91 – 99, (Morgan Kaufmann Publishers, San Francisco, 1998) ISBN: 1-55860-556-8 , http://research.microsoft.com/~fayyad

[12] Fayyad, U., Reina, C. and Bradley, P.S. Initialization of Iterative Refinement Clustering Algorithms, Proc. 4 th Int'l Conf. on Knowledge Discovery and Data Mining (KDD98), (AAAI press, 1998), http://citeseerx.ist.psu.edu/viewdoc/summary?doi=10.1.1.44.4060

[13] Hartigan, J. A. Clustering Algorithms (John Wiley & Sons, New York, 1975)

[14] Faber, V. Clustering and the Continuous k-means Algorithm, Los Alamos Science, 22, 138-144 (1994)

[15] Pena, J.M., Lozano, J.A. & Larranaga, P., An empirical comparison of four initialization methods for the K Means algorithm, Pattern Recognition Letters , 20 , 10, 1027 - 1040 (1999) ISSN:0167-8655

[16] Meila, M. and Heckerman, D. An experimental comparison of several clustering and initialization methods (1998), http://citeseerx.ist.psu.edu/viewdoc/summary?doi=10.1.1.8.1159

[17] Zhang, Z., Dai, B. T. and Tung, A. K.H. On the Lower Bound of Local Optimums in K-Means Algorithm, Proc. 6 th Intl Conf on Data Mining ICDM'06, 775 – 786 (IEEE Computer Society, Washington DC, 2006) ISBN: 0-7695-2701-7, DOI: 10. 1109/ ICDM.2006.118

[18] Daoud, M. A., Venkateswarlu, N. B. and Roberts, S. New Methods for the Initialization of Clusters, Pattern Recognition Letters 17, 5, 451 - 455 (1996) ISSN: 0167-8655