International Journal of Scientific Research in Computer Science, Engineering and Information Technology
© 2018 IJSRCSEIT | Volume 3 | Issue 6 | ISSN : 2456-3307

# Auto Determination of K in KMEANS with MAP-REDUCE for Numerical and Text Datasets

**Ms. K. P. Shiudkar[1*], Prof. S. B. Takmare[2], Prof. R. P.Mirajkar[3]**

[*1]ME CSE Student, Bharati Vidyapeeth College of Engineering, Kolhapur, Maharashtra, India
[2]Assistant Professor, Department of CSE, A P Shah Institute of Technology Thane, Maharashtra, India
[3]Assistant Professor, Department of CSE, Bharati Vidyapeeth college of Engineering Kolhapur, Maharashtra, India

## ABSTRACT

Data mining is the process of automatically discovering useful information in large datasets. Clustering analysis is a very important branch in data mining. Cluster analysis based on the data objects and their relationships and grouping of data objects. Clustering very large datasets is a challenging problem for data mining and processing. Map Reduce is considered as a powerful programming framework, which significantly reduces executing time by dividing a job into several tasks, and executes them in a distributed environment. K-Means, which is one of the most used clustering methods, and K-Means based on Map Reduce is considered as an advanced solution for very large dataset clustering. However, the executing time is still an obstacle due to the increasing number of iterations when there is an increase of dataset size and number of clusters. The traditional k-means is computationally expensive, sensitive to outliers and has an unstable result hence its inefficiency when dealing with very large datasets. Solving these issues is the subject of much recent research work. In this paper, we propose an Auto determination of K in KMEANS with MAP-REDUCE for numerical and text datasets in order to adapt it to handle large-scale datasets by reducing its execution time. In addition, we proposed algorithms to find the initial centroids automatically and cluster are formed on both numerical and text both datasets.

**Keywords :** Initial Centroids, Clustering, Data mining, Data sets, K-means clustering, Map-Reduce.

## I. INTRODUCTION

Big Data is evolving term that describes any voluminous amount of structured, semi-structured and unstructured data. Big data challenges include capturing data, data storage, data analysis, search, sharing, transfer, visualization, querying, updating, information privacy and data source. Big data represents the information assets characterized “5Vs”, volume (size of data set), variety (range of data type and source), velocity (speed of data in and out), value (how useful the data is), and veracity (quality of data) to require specific technology and analytical methods for its transformation into value. It creates challenges in their collection, processing, management and analysis. Big data to the use of predictive analytics, user behaviour analytics, or certain other advanced data analytics methods that extract value from data, and seldom to a particular size of data set. Big data analytics is the process of examining large and varied data sets to uncover hidden patterns, unknown correlations, market trends, customer preferences and other useful information that can help organizations make more-informed business decisions. As new data

and updates are constantly arriving, there is need of data mining to tackle challenges.

Data mining is the process of extraction of useful and interesting patterns from huge amount of data. It is called as knowledge discovery process and pattern analysis. Traditional data mining approach is not directly used for big data analysis. Data mining has various algorithms and techniques like Classification, Clustering, Regression, Artificial Intelligence, Neural Network, Association Rules, Decision trees, Genetic algorithms, Nearest Neighbour methods etc. are used for knowledge discovery from datasets.
Clustering is one of popular technique of grouping a set of objects in such a way that objects in one group(called cluster) are more similar to each other than to those other group. Several algorithms have been designed to perform clustering, each one uses different principle. They are divided into hierarchical, density-based, partitioning, grid-based and model based algorithms.

K-means is the most commonly used partitioning algorithm in cluster analysis because of its simplicity and performance. However, it has some restrictions when dealing with very large datasets because of high computational complexity, sensitive to outliers and its results depends on initial centroids, which are selected randomly. Many solutions have been proposed to improve the performance of KMeans. However, no one provide a global solution. Some of proposed algorithms are fast but they fail to maintain the quality of clusters. Some generate clusters of good quality but they are very expensive in term of computational complexity.

By taking all into consideration an enhanced Map-Reduce design used with k-means algorithm. Therefore, it will reduce the execution time while forming clusters on very large datasets. The outliers are major problem that will effect on quality of clusters. Some algorithm only works on numerical datasets.

### A. K-MEANs Clustering

K-means clustering technique is widely used clustering algorithm, which is most popular clustering algorithm that is used in scientific and industrial applications. It is a method of cluster analysis, which is used to partition N objects into k clusters in such a way that each object belongs to the cluster with the nearest mean [3].

The Traditional KMeans algorithm is very simple [3]:

1. Select the value of K i.e. Initial centroids.
2. Repeat step 3 and 4 for all data points in dataset.
3. Find the nearest point from those centroids in the Dataset.
4. Form K cluster by assigning each point to its closest centroid.
5. Calculate the new global centroid for each cluster.

Properties of k-means algorithm [3]:

1. Efficient while processing large data set.
2. It works only on numeric values.
3. The shapes of clusters are convex.

K-means is the most commonly used partitioning algorithm in cluster analysis because of its simplicity and performance. However, it has some restrictions when dealing with very large datasets because of high computational complexity, sensitive to outliers and its results depends on initial centroids, which are selected randomly. Many solutions have been proposed to improve the performance of KMeans. However, no one provide a global solution. Some of proposed algorithms are fast but they fail to maintain the quality of clusters. Some generate clusters of good quality but they are very expensive in term of computational complexity. The outliers are major problem that will effect on quality of clusters. Some algorithm only works on only numerical datasets.

### B. Review of Literature

Amira Boukhdhir , Oussama Lachiheb , Mohamed Salah Gouider [1] proposed algorithm an improved KMeans with Map Reduce design for very large dataset. The algorithm takes less execution time as compared to traditional KMeans, PKMeans and Fast KMeans. It removes the outlier from numerical datasets also Map Reduce technique used to select initial centroids and forming the clusters. But it has limitations like the value of numbers of centroids required as input by user. It works on numerical datasets only. In addition, numbers of clusters are not determined automatically.

Duong Van Hieu and Phayung Meesad [2] proposed algorithm for reducing executing time of the k-means .They implemented this by cutting off a number of last iterations. In the experiment method 30% of iterations are reduced, so 30% of executing time is reduced, and accuracy is high. However, the choosing randomly the initial centroids produces the instable clusters. Noise points may affect clustering result, so it produces inaccurate result.

Li Ma and al. [3] developed a solution for improving the quality of traditional k-means clusters. They used the technique of selecting systematically the value of k as well as the initial centroids. Also they reduced the number of noise points so the outlier's problem solved. This algorithm produces good quality clusters but it takes more computation time.

Xiaoli Cui and al. [4] proposed an algorithm an improved k-means. The algorithm is applied only to representative points instead of the whole dataset, used a sampling technique. The result of this the I/O cost and the network cost reduced because of Parallel K-means. Experimental results shows that the algorithm is efficient and it has better performance as compared with k-means but there is no high accuracy.

## II. METHODS AND MATERIAL

### A. Related Work

This approach aims in improving the performance in terms of reducing number of iteration by increasing number of data points and form the clusters by automatically finalizing the cluster centroids using Map-Reduce. Algorithm works on both numerical and text.

**Stage 1: Outlier removal**
**Stage 2: Auto determination of K**
**Stage 3: Cluster Formation**

Outlier removal removes the outliers (unnecessary data) from the datasets. For removing outliers first, we have to define the radius eps and value of number of neighbors (nb) for each data point. If numbers of neighbors are less than the defined neighbors (nb), then that data point declared as outlier. Remove that data point from dataset and repeat same process for all the data points in dataset.

In auto determination of K module, the number of clusters i.e. value of K is determined automatically. In this first, we use the classical step of KMeans with Map Reduce function to pick the initial cluster centers randomly. After running classical K-Means and defining, the initial centroid next step is to refined current centroids. The centroids are refined by choosing two new centres (C1 and C2) Then for each cluster centroid in K-Means; we picked the two new centroids. In addition, these two centroids used for next iterations. After finalizing the two new centers for each centroid, we tested the each cluster using Test cluster. Here the mapper checks the point of clusters on the line by joining the two centers (C1 and C2) and the reduce test for normal distribution of points on line.

In Cluster Formation, clusters are formed using the K centroids. These centroids are automatically selected

using Auto determination of K stage. The MAP-COMBINE-REDUCE functions are used for forming the clusters. There are two types of datasets numerical and text. Euclidian distance formula for forming clusters of numerical datasets. Cosine similarity formula for forming cluster of text datasets.

## III. RESULTS AND DISCUSSION

### A. Numerical Data Clustering

**Table 1.** Experimental Numeric Dataset

| Dataset (No.of Records) | No of Iterations |
|---|---|
| 1000 | 7 |
| 2000 | 7 |
| 3000 | 8 |
| 4000 | 9 |
| 5000 | 10 |

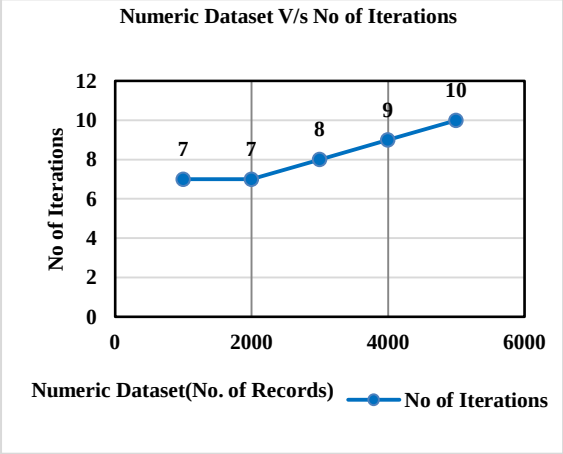


**Figure 1.** No of Numeric records/No of Iterations

In above experimental result shows the number of iterations vs. numerical dataset data point. Result analysis shows five numerical dataset with different size are chosen in the experiment and number of iterations is computed in the results. From Figure 4.1 we can see that the if the number of data points increased, it slightly increases the iterations. Even though the data points are increased twice, the iterations are not increased twice. Algorithm works efficiently.

### B. Text Clustering

**Table 2.** Experimental Text Dataset

| Dataset (No. of Documents) | No of Iterations |
|---|---|
| 50 | 12 |
| 100 | 13 |
| 150 | 14 |
| 200 | 15 |
| 250 | 16 |

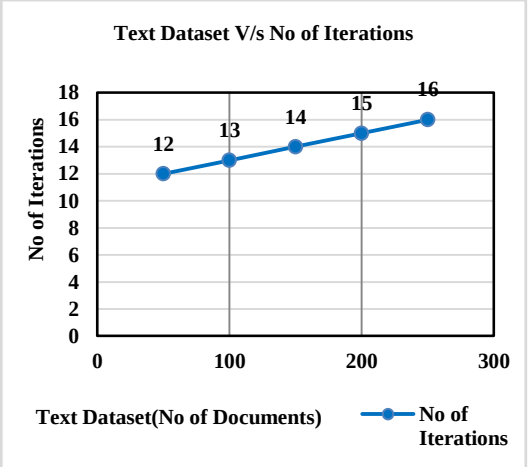


**Figure 2.** No of Text Documents / No of Iterations

In above experimental result shows the number of iterations vs. text dataset. Result analysis shows five-text dataset with different number of documents are chosen in the experiment and number of iterations is computed in the results. From Fig. 5, we can see that if the number of documents in the dataset increased, it slightly increases the iterations. Result obtained shows that a document increased twice does not increase the iterations twice. Algorithm works efficiently for text data set also.

### C. PERFORMANCE ANALYSIS OF MAP REDUCE ON CLUSTERS

To compute performance of K-Means with MAP REDUCE test performed by increasing the number of data points and analysis computed by observing the number of iterations. The performance analysis done on both numerical data and text data using determination of K clusters in KMeans with Map Reduce.

## IV. CONCLUSION

We have developed the module for Auto Determination of K in KMEANS with MAP-REDUCE for Numerical and Text Datasets. Experimental result shows first outliers are removed from datasets. Then the centroid of K-Means algorithm selected automatically using auto determination of K and the clusters are formed for numerical and text dataset using MAP-REDUCE. When number of data points for both numerical and text datasets are increased it will slightly effect on number of iterations. Algorithm works very efficiently and reduces execution time.

## V. REFERENCES

[1]. Amira Boukhdhir , Oussama Lachiheb , Mohamed Salah Gouider. "An Improved Map Reduce Design of Kmeans for clustering very large datasets", Published in IEEE/ACS 12th International Conference of Computer Systems and Applications (AICCSA)(2015)

[2]. V. Duon, M. Phayung. "Fast K-Means Clustering for very large datasets based on Map Reduce Combined with New Cutting Method (FMR. KMeans)", Springer International Publishing Switzerland, 2015.K. Elissa.

[3]. M. Li and al. "An improved k-means algorithm based on Map reduce and Grid", International Journal of Grid Distribution Computing, (2015)

[4]. C. Xiaoli and al. "Optimized big data K-means clustering using Map Reduce", Springer Science + Business Media New York (2014).

[5]. Thibault Debatty, Pietro Michiardi, Wim Mees, Olivier Thonnard, "Determining the K in KMEANS with Map Reduce" ,Published in the Workshop Proceedings of the EDBT/ICDT 2014 Joint Conference (March 28, 2014, Athens, Greece) on CEUR-WS.org (ISSN 1613-0073).

[6]. Document Clustering Using Improved K-Means Algorithm Shreyata khatri1,Dr. Kanwal Garg2 Research scholar,DCSA, Kurukshetra university,kurukshetra Assistant professor, DCSA Kurukshetra University, kurukshetra

[7]. K-means Clustering Optimization Algorithm Based on MapReduce Zhihua Li1,a, Xudong Song,b,WenhuiZhu,YanxiaChen,International symposium on Computers & Informatics (ISCI 2015)