# KNOWLEDGE CLUSTERING ON BIG DATA WITH K_MEANS ALGORITHM

S. Parthasarathy, V. Shakila

[1] *Assistant Professor(Senior Grade),Department of MCA, Valliammai Engineering College, Tamilnadu, India.*
[2] *2nd Year MCA Student, Department of MCA, Valliammai Engineering College, Tamilnadu, India.*

---

Abstract - *Big Data concerns with large-volume, complex, growing data sets in multiple and autonomous sources. Data storage and Data collection has become more complex. Big Data is now rapidly expanding in all science and engineering domains including physical, biological and biomedical sciences. This paper presents a k_means algorithm that characterizes the features of the Big Data revolution, and proposes a Big Data processing model from the Data mining perspective. This paper provides the most relevant and most accurate social sensing feedback to better understand our society at real time with big data technologies..*

*Key Words: Big Data, k_means, Data Mining...*

## 1. INTRODUCTION

Big Data starts with large volume and clustering using K_means algorithm. It seeks to explore more complex and evolving relationships among data. Huge data with heterogeneous and diverse dimensionality has autonomous sources with distributed and decentralized control in complex and evolving relationships. So the noise is one of the problems. In this paper, we group the objects by using clustering which is similar to one cluster and is dissimilar group in to other cluster. To overcome this problem, we use k_means algorithm by implementing using java programming language.

### 1.1 Big Data

The most fundamental challenge for Big Data applications is to explore the large volumes of data and extract useful information or knowledge for future actions. In many situations, the knowledge extraction process has to be very efficient and close to real time because storing all observed data is nearly infeasible. Sometimes large amount of data is beyond the software tools to manage.

### 1.2 Noise and Error

In other situation, privacy concerns with noise and errors. These can be interested into the data to produce altered data copies. At the data level, the autonomous information sources and the variety of the data collection environments often result in data with complicated conditions such as missing or uncertain values.

## 2. Theories and Challenges

System analysis is the process of examining an existing system of methodology with intend of improving it through better procedure and methods after analyzing the requirements of the task to be performed. The next step is to analyze the problem and understand its context. The first activity in the phase is studying the existing system and other is to understand the requirements and domain of the new system. Both the activities are equally important, but the first activity serves as a basis of giving the functional specifications and then successful design of the proposed system.

### 2.1 HACE Theorem

The Existing system presents a HACE theorem that characterizes the features of the Big Data revolution, and proposes a Big Data processing model, from the data mining perspective. This data-driven model involves demand-driven aggregation of information sources, mining and analysis, user interest modeling, and security and privacy considerations.

### 2.2 Challenges of Tier Architecture

The challenges at Tier-I focus on data accessing and arithmetic computing procedures. The challenges at Tier-II center on semantics and domain knowledge for different Big Data applications. At Tier-III, the data mining challenges concentrate on algorithm designs in tackling the difficulties raised by the Big Data volumes, distributed data distributions, and by complex and dynamic data characteristics.

### 2.3 k_means Algorithm

The proposed system based on clustering algorithm. Clustering is a collection of objects which are “similar” between them and are “dissimilar” to the objects belonging to other cluster by k_means algorithm. It is one the exclusive clustering algorithm and is one of the simplest unsupervised learning algorithm that solve the well known clustering problem, which is used to group the similar objects in one cluster and dissimilar objects in

other cluster. Preliminary investigation examines the project feasibility, the likelihood how the system will be useful to the organization. The main objective of the feasibility study is to test the technical, operational and economical feasibility for adding new modules and debugging old running system. All system is feasible if they are unlimited resources and infinite time. There are aspects in the feasibility study portion of the preliminary investigations on operational, technical and economical feasibility.

### 2.3 Clustering with Big Data

K means algorithm is one of the clustering algorithms. Clustering is a collection of objects which are "similar" between them and are "dissimilar" to the objects belonging to other cluster. By using java program the user gets the input based on that it performs operation and group the objects which is used for the analysis task for the project. The technique which is used in this paper known as Clustering concepts in data mining to avoid the problem of the noisy data in the big data project. Every day, 2.5 quintillion bytes of data are created and 90 percent of the data in the world today were produced within the past two years. The world will generate 50 times the amount of information and 75 times the number of information contains by 200 while information Technology personnel to manage. It will grow less than 1.5 times.

## 3. System Architecture and Design

In System Architecture, the user wants to gather data from the any data source then preprocess the data. Then it stores in the dataset which here act as MySQL database. It uses the technique called clustering with k_means algorithm and gets the extract data or knowledge for the future analysis of task. Therefore this algorithm is mostly useful for the user to get the extract information while having the large amount of data.

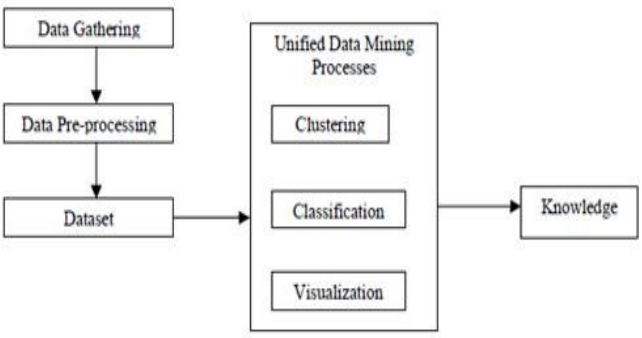



Fig -1: System Architecture

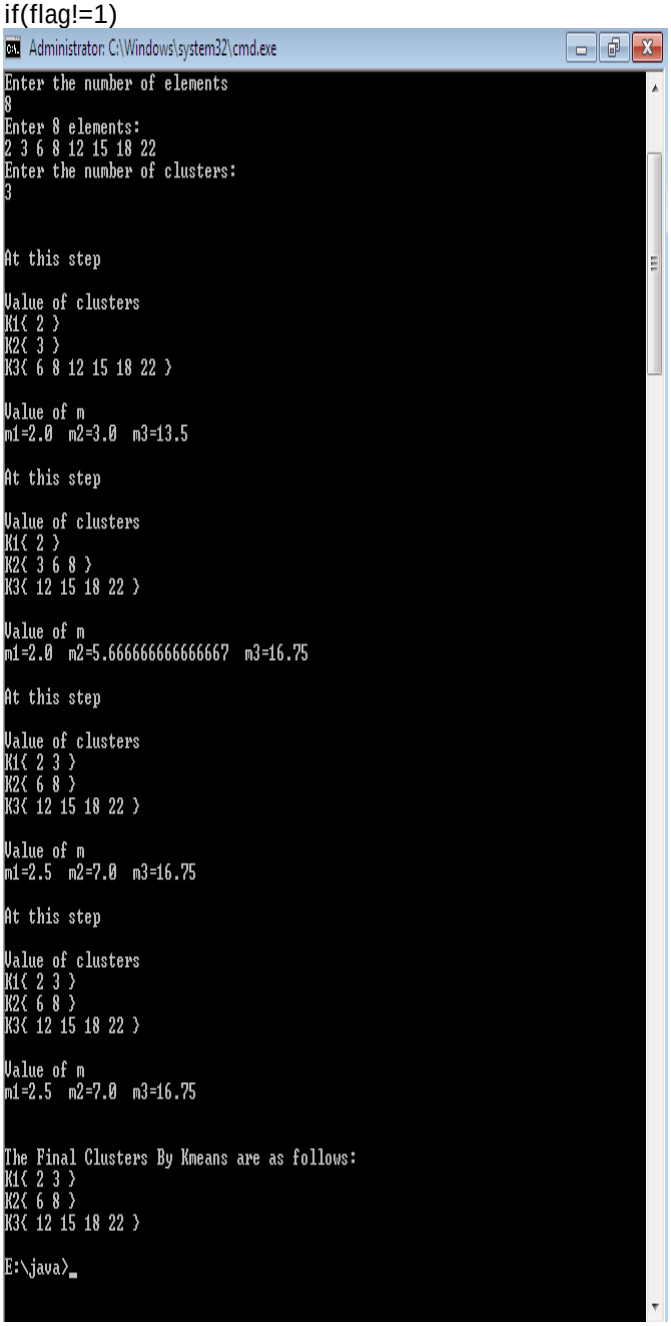
## 4. System Implementation

In the implementation program, we take eight numbers of elements then the number of cluster is three, so that it will calculate the mean value between neighbor values and place in three clusters. This is the process of K_means algorithm. Finally we get the values for further analysis using clustering in data mining. So this will help us to process the big data without any noisy data for the project. Then the user wants to gather data from the any data source then preprocess the data which gathered by the user, then store it in the dataset which here act as MySQL database. By using the technique called clustering with k_means algorithm gets the extract data or knowledge for the future analysis task.

```
import java.util.*;
class k_means
{
static int count1,count2,count3;
static int d[];
static int k[][];
static int tempk[][];
static double m[];
static double diff[];
static int n,p;

static int cal_diff(int a) // This method will determine the
cluster in which an element go at a particular step.
{
int temp1=0;
for(int i=0;i<p;++i)
{
if(a>m[i])
diff[i]=a-m[i];
else
diff[i]=m[i]-a;
}
int val=0;
double temp=diff[0];
for(int i=0;i<p;++i)
{
if(diff[i]<temp)
{
temp=diff[i];
val=i;
}
}//end of for loop
return val;
}

static void cal_mean() // This method will determine
intermediate mean values
{
for(int i=0;i<p;++i)
m[i]=0; // initializing means to 0
int cnt=0;
for(int i=0;i<p;++i)
{
cnt=0;
for(int j=0;j<n-1;++j)
```

```
{
if(k[i][j]!=-1)
{
m[i]+=k[i][j];
++cnt;
}}
m[i]=m[i]/cnt;
}
}
static int check1() // This checks if previous k ie. tempk and current k are same.Used as terminating case.
{
for(int i=0;i<p;++i)
for(int j=0;j<n;++j)
if(tempk[i][j]!=k[i][j])
{
return 0;
}
return 1;
}

public static void main(String args[])
{
Scanner scr=new Scanner(System.in);
/* Accepting number of elements */
System.out.println("Enter the number of elements ");
n=scr.nextInt();
d=new int[n];
/* Accepting elements */
System.out.println("Enter "+n+" elements: ");
for(int i=0;i<n;++i)
d[i]=scr.nextInt();
/* Accepting num of clusters */
System.out.println("Enter the number of clusters: ");
p=scr.nextInt();
/* Initialising arrays */
k=new int[p][n];
tempk=new int[p][n];
m=new double[p];
diff=new double[p];
/* Initializing m */
for(int i=0;i<p;++i)
m[i]=d[i];

int temp=0;
int flag=0;
do
{
for(int i=0;i<p;++i)
for(int j=0;j<n;++j)
{
k[i][j]=-1;
}
for(int i=0;i<n;++i) // for loop will cal cal_diff(int) for every element.
{
temp=cal_diff(d[i]);
if(temp==0)
k[temp][count1++]=d[i];
else
if(temp==1)
k[temp][count2++]=d[i];
else
if(temp==2)
k[temp][count3++]=d[i];
}
cal_mean(); // call to method which will calculate mean at this step.
flag=check1(); // check if terminating condition is satisfied.
if(flag!=1)
```



## 3. CONCLUSIONS

We regard Big Data as an emerging trend and the need for big data mining is arising in all Science and Engineering domains. With big data technologies, we will hopefully be able to provide most relevant and most accurate social sensing feedback to better understand our society at real time. We can further stimulate the participation of the public audiences in the data production circle for societal and economical events. The era of big data has arrived.

## REFERENCES

[1] R. Ahmed and G. Karypis, "Algorithms for Mining the Evolution of Conserved Relational States in Dynamic Networks," Knowledge and Information Systems, vol. 33, no. 3, pp. 603-630, Dec. 2012S.

[2] M.H. Alam, J.W. Ha, and S.K. Lee, "Novel Approaches to Crawling Important Pages Early," Knowledge and Information Systems, vol. 33, no. 3, pp 707-734, Dec. 2012.

[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. 19, no. 1, pp. 20-23, 2012.

[5] S. Banerjee and N. Agarwal, "Analyzing Collective Behavior from Blogs Using Swarm Intelligence," Knowledge and Information Systems, vol. 33, no. 3, pp. 523-547, Dec. 2012

## BIOGRAPHIES

S. Parthasarathy (Sellaperumal Parthasarathy) obtained his Bachelor's degree in Mathematics, B.Ed. degree in Education, Master's degree M.C.A, M.Phil. Degree from the department of Computer Science, M.Tech. Degree from the department of Computer Science and Engineering, Professional M.B.A. Degree and perusing Ph.D in Computer Science. He has also obtained EMC Academic Associate in Cloud Infrastructure and Services. Currently, he is an Assistant Professor at the Faculty of MCA, Valliammai Engineering College under Anna University. His specializations include Scheduling in Cloud Computing.

V. Shakila, pursued her Bachelor of Computer Application degree from Prince Shri Venkateshwara Arts and Science College under Madras University. Currently, She is doing her Master of Computer Application in Vallimmai Engineering College under Anna University. Her area of interest is Data Mining.