ISSN No: 2348-4845
International Journal & Magazine of Engineering, Technology, Management and Research
*A Peer Reviewed Open Access International Journal*

# Clustering in Big Data Using K-Means Algorithm

Ajitesh Janaswamy, B.E
**Dept of CSE,**
**BITS Pilani, Dubai Campus.**

## ABSTRACT:

K-means is the most widely used clustering algorithm due to its fairly straightforward implementations in various problems. Meanwhile, when the number of clusters increases, the number of iterations also tends to slightly increase. In this study, improved implementations of k-means algorithm with a centroid calculation heuristics which results in a performance improvement over traditional k-means are proposed. The results show that big data implementation model outperforms the other compared methods after a certain threshold level and small data implementation performs better with increasing k value.

## INTRODUCTION:

The rate of data creation at present has increased so much that 90% of the data in the world today has been created in the last two years alone. This huge amount of data is being viewed by business organizations and researchers as a great potential resource of knowledge that needs to be discovered. Traditional methods of data analysis and management do not suffice. New technologies to deal with this data called Big Data are required. The term Big Data refers to large-scale information management and analysis technologies that exceed the capability of traditional data processing technologies.

The incorporation of Big Data is changing Business Intelligence and Analytics by providing new tools and opportunities for leveraging large quantities of structured and unstructured data. Big Data is notable not because of its size, but because of its relationality to other data. Due to the methods used to store the data, Big Data is fundamentally networked (threaded with connections). But these connections are not useful directly. The actual value comes from the patterns that can be derived from the related pieces of data about an individual, about individuals in relation to others, about groups of people, or simply about the structure of information itself. Besides this, Big Data has enormous volume, high velocity, much variety and variation. These features of Big Data present the main challenges in analyzing Big data which are: (1) Efficient and effective handling of large data, (2) Processing time and accuracy of results trade –off; and (3) Filtering important and relevant data from all the data collected. Traditionally, large data is handled through numerous data mining techniques. Recognizing patterns among data must borrow ideas from Machine learning algorithms. Thus, Big data analysis needs fusion of techniques for data mining with those of machine learning. The k-means algorithm is one such algorithm which has presence in both the fields.

K-means is one of the most famous partition clustering algorithms because it is a very simple, statistical and quite scalable method. Also it has linear asymptotic running time with respect to any variable of the problem. Yet, k-means cannot be used for Big Data analysis directly. It needs to be adapted to deal with sparse values, heterogeneity and velocity. This paper emphasizes the need of filtering data before it is analyzed for information. The strategic information, that the Business Analysts seek, always has some defined line of interest. Filtering of data should reflect this so that results of analysis can be of value to the analysts.

Also, given the high velocity of Big Data, the research should be directed towards approximate and heuristic solutions for clustering instead of ideal ones. Finding ideal clusters consumes time which in some cases may render the information deduced to be stale. Hence, an approximate algorithm which reduces the complexity of classic k-means by computing over only those attributes which are of interest is proposed here.

The drawbacks of traditional clustering algorithms have been identified and the proposed solution is an effort to overcome them.

### Existing System:

Clustering**,** or cluster analysis, is an important subject in data mining. It aims at partitioning a dataset into some groups, often referred to as clusters, such that data points in the same cluster are more similar to each other than to those in other clusters. There are different types of clustering. However, most of the clustering algorithms were designed for discovering patterns in static data. This imposes additional requirements to traditional clustering algorithms to rapidly process and summarize the massive amount of continuously arriving data.

### Disadvantages:

It also requires the ability to adapt to changes in the data distribution, the ability to detect emerging clusters and distinguish them from outliers in the data, and the ability to merge old clusters or discard expired ones.

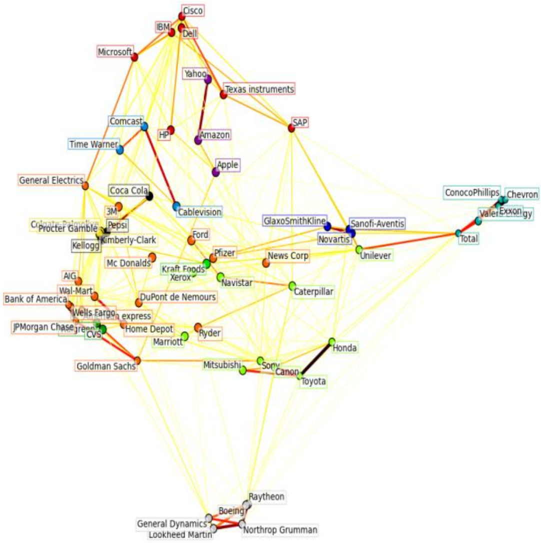
### Proposed System:

We extend a recently proposed clustering algorithm, affinity propagation (K-Means) clustering, to handle dynamic data. Several experiments have shown its consistent superiority over the previous algorithms in static data. K-Means clustering is an exemplar-based method that realized by assigning each data point to its nearest exemplar, where exemplars are identified by passing messages on bipartite graph. There are two kinds of messages passing on bipartite graph. They are responsibility and availability, collectively called 'affinity'.K-Means clustering can be seen as an application of belief propagation, which was invented by Pearl to handle inference problems on probability graph. Compared with the previous works, another remarkable feature of our work is that the K-Means clustering algorithms are proposed based on a message-passing framework. That's, each object is a node in a graph, and weighted edges between nodes correspond to pair wise similarity between objects. When a new object is observed, it will be added on the graph and then message passing is implemented to find a new exemplar set. Because that only one, or a few of nodes' entering will not change the structure of the whole graph a lot, a local adjustment of availabilities and responsibilities is enough. Therefore, messages passing on graphs will re-converge quickly. Based on these features, the K-Means clustering algorithms proposed in this paper don't need to re-implemented K-Means clustering on the whole data set, nor need to change the similarities between objects.

### Advantages:

1. A great deal of time can be saved, which makes K-Means clustering efficient enough to be used in dynamic environment.
2. The goal of this paper is to propose a dynamic variant of K-Means clustering, which can achieve comparable clustering performance with traditional K-Means clustering by just adjusting the current clustering results according to new arriving objects, rather than re-implemented K-Means clustering on the whole dataset.

## SYSTEM ARCHITECTURE:



## IMPLEMENTATION

### Modules:

1. K-Means Clustering

2. K-Means Clustering Based on K-Medoids
3. K-Means Clustering Based on Nearest Neighbor Assignment:
4. Test by Labeled Data Sets

### 1. K-Means Clustering:

Clustering is one of the most important clustering algorithms. It is realized by firstly picking out some special objects that called exemplars, and then associating each left object to its nearest exemplar. The objective is to maximize

$$z = \sum_{i=1}^{n} s(i, c_i),$$

where s(i; ci) denotes similarity between **x**i and its nearest exemplar **x**ci . The most attracting advantage of exemplar-based clustering is that the exemplar set itself stores compressed information of the whole data set. However, find an optimal set of exemplars is essentially a hard combinational optimization problem.

### 2. K-Means Clustering Based on K-Medoids:

K-Means clustering has been successfully used in a series of problems, e.g., face recognition, f MRI data analysis, and document clustering. However, most of the applications deal with static data. Incremental AP clustering is still a difficult problem. The difficulty in incremental AP clustering is that: after affinity propagation, the first batch of objects have established certain relationships (nonzero responsibilities and nonzero availabilities) between each other, while new 'objects' relationships with other objects are still at the initial level (zero responsibilities and zero availabilities). Objects arriving at different time step are at the different statuses, so it is not likely to find the correct exemplar set by simply continuing affinity propagation in this case. Fig. 2 is a toy example to demonstrate such a problem, where the data come from the computational experiments. Traditional AP clustering is implemented on the first batch of objects. Responsibilities and availabilities converge.

### 3. K-Means Clustering Based on Nearest Neighbor Assignment:

The K-Means clustering algorithm is proposed according to the second strategy. A technique of Nearest-neighbor Assignment (NA) is employed to construct the relationships (values of responsibilities and availabilities) between the new arriving objects and the previous objects. NA means that the responsibilities and availabilities of the new arriving objects should be assigned referring to their nearest neighbors. NA is proposed based on such a fact that if two objects are similar, they should not only be clustered into the same group,but also have the same relationships (responsibilities and availabilities). However, most of the current algorithms utilize the former part only.

### 4. Test by Labeled Data Sets:

The clustering performance of an algorithm is evaluated by external dispersity and internal dispersity. The sum of similarities is one of the most widely used criteria. In some cases, different clustering result can obtain comparable external dispersity and internal dispersity. Therefore, we use labeled data sets to evaluate the proposed algorithms in this section. An advantage is that we can not only evaluate the clustering algorithms by dispersity, but also by some other indicators, e.g., mutual information, clustering accuracy. According to the object function of exemplar-based clustering in (1), the Sum of Similarities (SS) is defined as

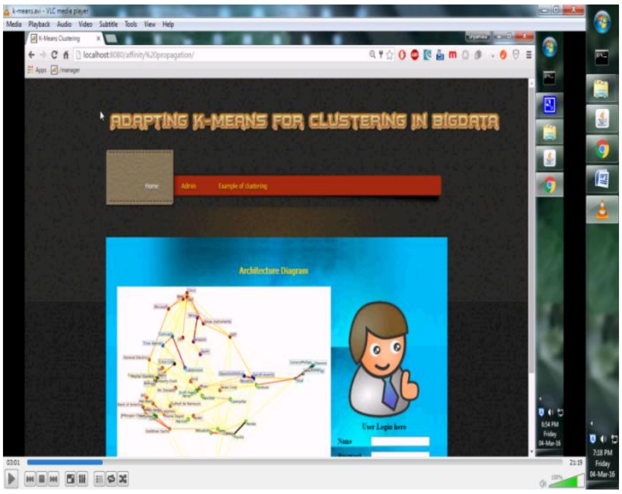
$$SS = \sum_{i=1}^{n} s(i, c_i).$$

A larger SS indicates a better clustering performance.

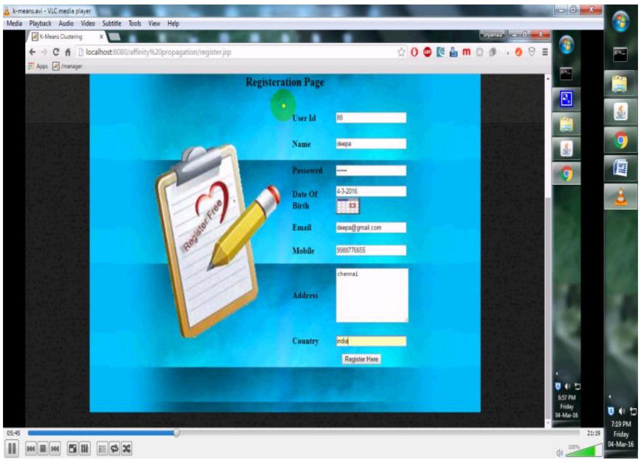
### Euclidean distance:

$$s(i, j) = -\sqrt{\| \mathbf{x}_i - \mathbf{x}_j \|^2}.$$

### Home page:


k-means.avi - VLC media player
localhost:8080/affinity%20propagation/
ADAPTING K-MEANS FOR CLUSTERING IN BIGDATA
Home
Admin
Example of clustering
Architecture Diagram
User Login here
Name
Password

### User Registration page:


k-means.avi - VLC media player
localhost:8080/affinity%20propagation/register.jsp
Register Free

Registeration Page

User Id 88

Name deepa

Passowrd •••••

Date Of Birth 4-3-2016

Email deepa@gmail.com

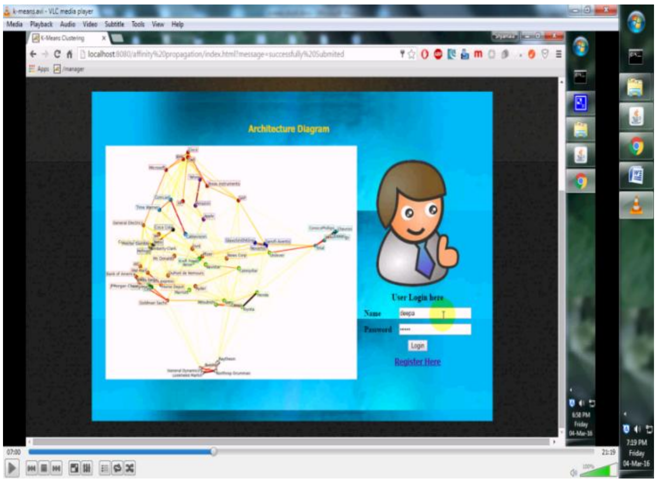
Mobile 9988776655

Address chennai

Country india

Register Here

### User Login page:


k-means.avi - VLC media player
localhost:8080/affinity%20propagation/index.html?message=successfully%20submitted
Architecture Diagram

User Login here

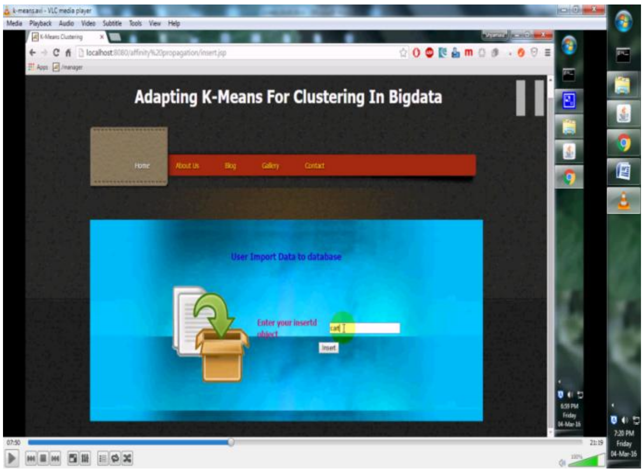
Name deepa

Password ••••

Login

Register Here

### User Import the data into database:


k-means.avi - VLC media player
localhost:8080/affinity%20propagation/insert.jsp
Home
About Us
Blog
Gallery
Contact

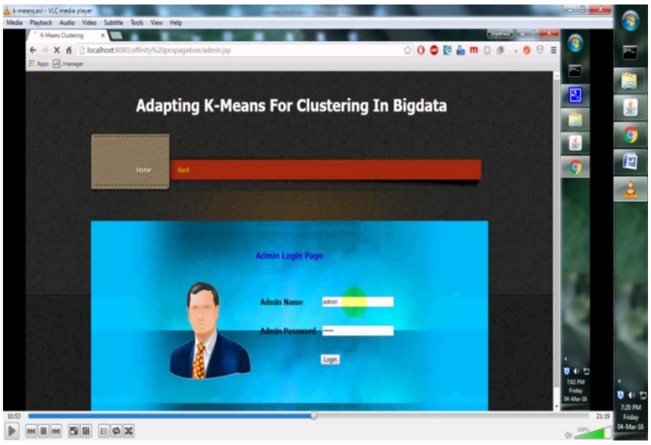
Adapting K-Means For Clustering In Bigdata

User Import Data to database

Enter your inserted object car

Insert

### Admin login:


k-means.avi - VLC media player
localhost:8080/affinity%20propagation/admin.jsp
Home
Back

Adapting K-Means For Clustering In Bigdata

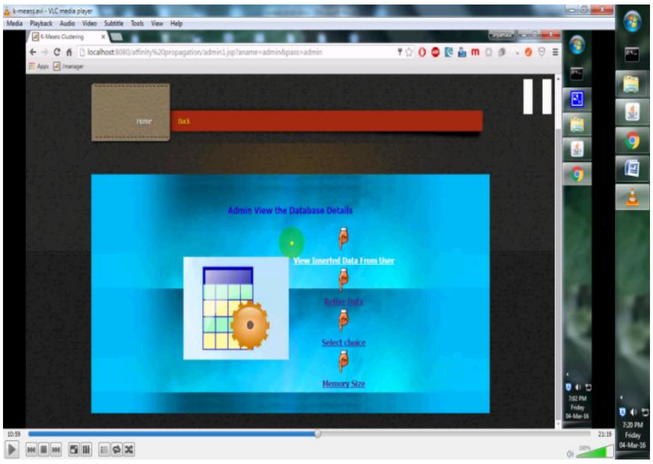
Admin Login Page

Admin Name admin

Admin Password •••••

Login

### Admin view page:


k-means.avi - VLC media player
localhost:8080/affinity%20propagation/admin1.jsp?aname=admin&apass=admin
Home
Back

Admin View the Database Details

View Inserted Data From User

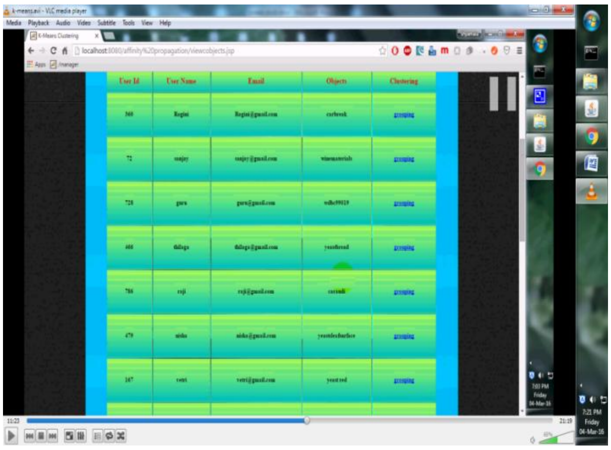
Refine Data

Select choice

Memory Size

ISSN No: 2348-4845
International Journal & Magazine of Engineering, Technology, Management and Research
*A Peer Reviewed Open Access International Journal*

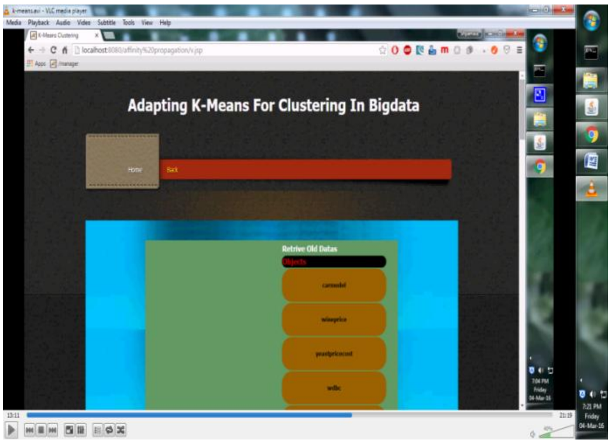
### Waiting for clustering objects:



localhost:8080/affinity%20propagation/viewobjects.jsp

| User Id | User Name | Email | Objects | Clustering |
|---|---|---|---|---|
| 360 | Regini | Regini@gmail.com | carbreak | grouping |
| 72 | sanjay | sanjay@gmail.com | winematerials | grouping |
| 728 | guru | guru@gmail.com | wdbc99019 | grouping |
| 466 | dhilaga | dhilaga@gmail.com | yeastbread | grouping |
| 786 | raji | raji@gmail.com | carindi | grouping |
| 479 | nisha | nisha@gmail.com | yeastsurface | grouping |
| 167 | vetri | vetri@gmail.com | yeast red | grouping |

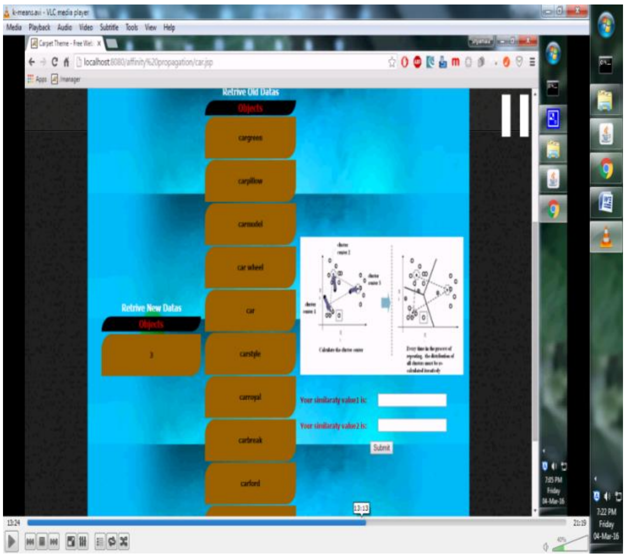
### View the overall objects:



localhost:8080/affinity%20propagation/v.jsp

Adapting K-Means For Clustering In Bigdata

### Clustering model:



localhost:8080/affinity%20propagation/car.jsp

Your similaraty value1 is:

Your similaraty value2 is:

### CONCLUSION:

Big Data is being viewed by all, from scientists to businesses, as potential resource of information. The information is not directly available and needs to be extracted from Big Data. Existing technologies are insufficient to be deployed for big data analysis. Formal architectures, new algorithms and fast heuristics to deal with the challenges posed by big data like volume, velocity and variety is the need of hour. In this paper an approximate method based on the classic k-means algorithm is suggested. The achievement is lowered time complexity and fixed number of iterations which depend only on the number of attributes to handle. Manhattan distance concept in a modified form has been used, which in turn decreases the run time. The efficacy and precision of algorithm is demonstrated on various real and synthetic datasets.

For most of the datasets, the precision achieved by the proposed algorithm is higher than the k-means and other contemporary popular clustering algorithms. Cluster recovery is also higher than most of them since the proposed algorithm does not reject any data. The concept of hierarchical clustering can be used along with the proposed algorithm to handle very large number of dimensions. Proposed work can be modified for rejecting flash data. The algorithm presented here cannot handle categorical data well until it is converted into equivalent numerical data. Exploring clustering big data in terms of categorical data could be another possible extension. Deciding primary and secondary attributes is considered in the proposal to be provided as an input by the user.

### REFERENCES:

http://www-01.ibm.com/software/data/bigdata/

Gartner's Big Data Definition Consists of Three Parts, Not to Be Confused with Three "V"s, By Svetlana Sicular, Gartner, Inc. 27 March 2013. [online] http://www.forbes.com/sites/gartnergroup/2013/03/27/ gartners-big-datadefinition-consists-of-three- parts-not-to-be-confused-with-three-vs/.

Italiano G.F. Algorithms for Big Data: Graphs and Memory errors. July 2013. Available online at almada2013.ru/files/courses/italiano/00-Intro.pdf

Forgy, E. W. Cluster analysis of multivariate data: Efficiency versus interpretability of classifications. Biometrics, 21:768–780, 1965.

MacQueen, J.B. Some methods for classification and analysis of multivariate observations. In Proc. of 5th Berkeley Symposium on Mathematical Statistics and Probability, volume 1, pages 281–297, 1967.

Lloyd, S. P. Least squares quantization in PCM. IEEE Transactions on Information Theory, 28(2):129–137, March 1982.

Extracting Value from Chaos, By Gantz, J. and Reinsel, D. IDC IVIEW June 2011. [online] http://www.emc.com/collateral/analyst-reports/idc-extracting-value-from-chaos-ar.pdf.

The Big Data Long Tail. Blog post by Bloomberg, Jason. On January 17, 2013. [online] http://www.devx.com/blog/the-big-data-long-tail.html.

The Fourth Paradigm: Data-Intensive Scientific Discovery. Edited by Hey, T, Tansley, S. and Tolle, K.. Microsoft Corporation, October 2009. ISBN 978-0-9825442-0-4.

Demchenko, Y., Membrey, P., Grosso, C. de Laat, Addressing Big Data Issues in Scientific Data Infrastructure. First International Symposium on Big Data and Data Analytics in Collaboration (BDDAC 2013). Part of The 2013 Int. Conf. on Collaboration Technologies and Systems (CTS 2013), May 20-24, 2013, San Diego, California, USA.
Ng, R. T., and. Han, J. Clarans: A method for clustering objects for spatial data mining. IEEE Transactions on Knowledge and Data Engineering (TKDE), 14(5):1003–1016, 2002.

Bezdek, J. C., Ehrlich, R., and Full, W. Fcm: The fuzzy c-means clustering algorithm. Computers & Geosciences, 10(2):191–203, 1984.

Zhang, T., Ramakrishnan, R., and Livny, M. Birch: an efficient data clustering method for very large databases. ACM SIGMOD Record, volume 25, pp. 103–114, 1996.

Hinneburg, A., and Keim, D. A. An efficient approach to clustering in large multimedia databases with noise. Proc. of the ACM SIGKDD Conference on Knowledge Discovery ad Data Mining (KDD), pp. 58-65, 1998.

Hinneburg, A., and Keim, D. A. Optimal Grid-clustering: Towards breaking the curse of dimensionality in high-dimensional clustering. In Proceedings of the 25th Conference on VLDB, 506-517, 1999.