International Academic Institute for Science and Technology

International Academic Journal of Science and Engineering
Vol. 2, No. 2, 2015, pp. 182-192.
ISSN 2454-3896

International Academic Journal of Science and Engineering

www.iaiest.com

Clustering Algorithms in Big data

Shirin Abbasi[a] , Babak Vaziri[b]

[a] *Master Student, Islamic Azad University, Centeral Tehran Branch,tehran, Iran.*

[b] *Assistant Professor,Islamic Azad University,Centeral Tehran Branch,Tehran, Iran.*

Abstract

Massive data processing and management is taken into consideration in recent years, frameworks and special tools have been introduced for this purpose and data analyzing is studied as massive data management. Data mining and machine learning techniques, in many organizations and industries are used to improve business, forecasting and boom And according to enterprise's data growth and high-volume and high-speed data generation, data mining techniques in massive data, play an important role in these developments. Operations carried out on data management that group of such operation are as preprocessing which is performed on data to prepare data for basic operations. On the other hand in cloud environments, due to environment distribution, grouping and data sorting operation is important greatly. In this paper we have studied the main clustering algorithm K-Means, and due to studies on this algorithm, some ways suggested that consider main parameters in cloud environments and improve the massive data clustering.

Keywords: Clustering – massive data- massive data management – cloud storage – cloud computing – K-Means

1-Introduction:

In the last decade with the spread of mobile phones and connecting to the web and also extention of sensors, data have been produced by high volume which have different dimensions and must be processed for different operation.
For example in social web sites such as face book,500 terabytes data have been produced daily
Big data are group of data which stored structured or non –structured and are complex data which has different dimensions.
The first feature of big data is volume which refer to their quantity and due to their large volume analyze of them is different and cannot be through traditional database .
If a small number of processing nodes to be used, due to the high volume, low speed processing takes place.
To speed up processing, more nodes and more processing power is needed that require higher fees.
Data compression is one of the principal ways which is suggested in this regards. It does not work on big data because one of the characteristics of big data, is their diversity.

These data combined of different types such as movie, photo, text, etc, Which this non-structuring, has been caused compression with difficulties.
And in some situation, spent time for processing data with traditional way, waste for compression, and on the other hand due to this various type, it has its own complexity .For this reason, compression is not applicable in big data processing ,
The next item which must be considered in big data processing is that these data are used in applications that transfer data on line or must be analyzed in a situation that can respond users in specific time, so in processing timing, timing is such that respond immediately.

Traditional data management is not applicable for big data management.
In data management all items such as different data structure, different dimension and their non-structural must be considered. [Grolinger,2013]

Special cases must be considered in big data management:

1.If Parallel processing is used, Nodes in the cluster must act in such a way that in the event of failure of any particular node, avoid any disruption in the processing of big data

2.System files that are used for big data must cover large volume of big data and their capacity must be GB or higher

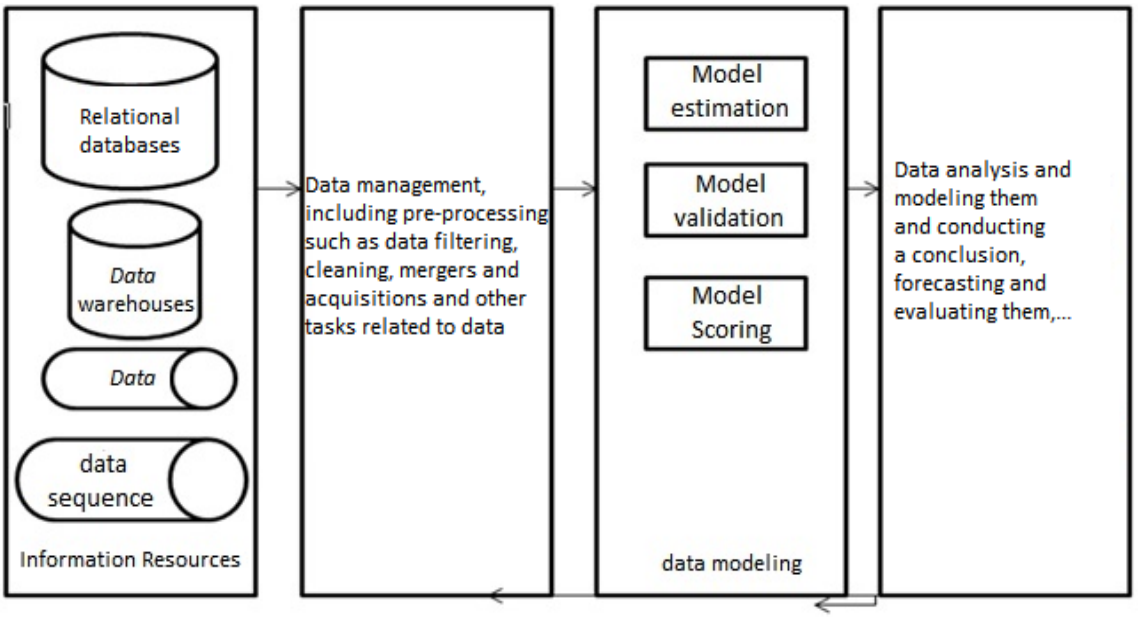
3.Reading and writing operations in many applications, run sequentially , and must be optimized for better high speed responding.

4.Some applications for better responding must be transferred to applicable program and to get rid of hardware compatibility and communication needs, it's better to use cloudy environments to cover this.
While big data discussion arises, Cloud computing is greatly raised, because there are common concerns among both problems, In both hardware and software resources are in a same network
And for users, Physical sub-structure is simulated that based on their requirement, and unlimited space is allocated to them apparently.
Cloudy database, is used for Appropriate allocation of resources in big data and because of parameters such as scalability and availability to be mentioned as a challenge a for cloudy environment

2-Previous works

In years of 1970decades, the basic concepts of the database proposed
While data volume increased gradually, using one computer for processing was impossible
At first used them for data storage, and then with expanding of relational data concept, for data analyzing, search and researches methods in data was expanded that data processing possibility provided.
In the early '80s, Take advantage of multiple systems and data sharing were provided
And each system has its own processor and memory and performs processing independently.
Due to the increasing volume of data related and need of process them in same time, different ways for big data has been considered.
While google faced to problems of low memory and analyze of many websites, expanded the file of systems which were using and used analyze and distributed process model
With the parallel processing, google designed high scalable database which named BigTable ,This type of database was able to Indexing and labeling data, so because of this ,data accessibility was done more quickly, and regarding using in google products and services, it arises as start point for big data processing and after cloud environments extending ,it was used as a potential standard in big data processing and search field.
Google technology wasn't open text technology and because of this Yahoo for extending this database, acts based on open text frameworks such as Hadoop which Hbase and Hive designed based on this.
In the management of Big Data there are different approaches , such as using big Table, Non-relational data management procedures [Kossmann et al,2010],or data management as existence ,or attribute Creation based on data different dimensions that each of them can provided only a part of needed characteristics of cloud environments.
In this research with Parameters extraction, an approach will be used based on combination of these cases which can covers cloudy environments characteristics, also in current cloudy storage services such as Amazon ,profit by approaches which based on them capability of insurance about stored data increased highly ,and these services are used for saved data. Such of these services is Dyamo [Chung,2008].
Services which is used will be studied And on the intersection of these two search, a solution for big data management in cloudy environment will be presented, and based on presented solution
A Search algorithm based on non relational database proposed for more coverage of big data and more success in responding, which is usable in normal data concepts, NOSQL.
The method by comparing between different search methods in the query languages that are used in large data will be achieved, also for better coverage of cloudy environments, also Compatibility level in this method will be evaluated.
Talia has researched about data complexity and variety and usage of cloudy computing as suitable frame work for processing of big data, And about allocation of resources with high need of big data processing, proposed approaches that used in special situation of cloudy environments, but cannot cover all needed characteristics by big data.
And most of these approaches are in such way that Map /reduce steps are accordingly arranged
But accessibility is not guaranteed in them and caused problem in cloudy environments [Guo et al,2010]
In most of offered articles in this regard, one needed characteristics of cloudy computing is considered, not all needed specifications and requirement of data management
If we consider three main characteristics, data integration, accessibility and flexibility in cloudy environment, and study them based on CAP thesis, that believes software systems in each case can cover just two of these main items, comparing and relational models can support just integration and accessibility
Key(clue) based models, just cover accessibility and divisibility, and models based Bigtable cover integration and divisibility and models based on documents cover accessibility and divisibility

If we want to study ACID characteristics in these approaches, again these approaches support a group of these characteristics.
As mentioned , NOSQL concepts, doesn't cover these scales and whole data integrity cannot been guaranteed[Vogels ,2009] but because of high volume of data in big data, in many cases , this characteristic has been ignored , but in cloudy environments , these characteristics are very important and in this research with derivation of service parameters in cloudy environments, we will make more compatible data managements approaches .
In article [Zadrozny et al,2013], a frame work has been offered as workflow which can consider it as reference Model for describing data managements approaches.
Big data will be obtained from different sources and due to high volume of them and their variety, it's necessary that pre processing such as filtering, data integration, cleanup information have been done, that data transfer to processing step.
After processing data and Placement in memory, they become as model and based on data models, assessment and forecasting operations are carried out, in below figure this procedure has been described:



Big Data management framework

Many solutions has been offered for analyzing big data, which also has been used in cloudy environments.
First solutions which proposed in this regards is system files using in internet, which Google and Yahoo have used them that can provide scalability and reliability in internet services.[Decandia et al,2007]
Also some models offered which emphasizes on existence of data centers in multiple Geographical regions, same as Amazon Cloudy storage service.
This time, parallel processing raised which based on map / reduce and also HADOOP workflow frame work was formed.

And then for better responding to high volume of queries same as social networks queries, query language comes at high level, which for first time in facebook,Hive was used at the same time with Hadoop for improving system performance.
The work frame which has been used for this case was such that can compress data due to high volume of them[stonebreaker et al,2007]
In this regard, some frame works had been offered which can be match to need of cloud environments, including usage of language and NOSQL frames works which provides possibility of non-structured unrelated data management.
To cover things like integrity, the accessibility and scalability, in many environments, these tools and methods are used in combination
Clustering information is a key issue in the various fields of computer science that at the various issues to be addressed, also in many of the methods and strategies used to label and classify them.[Ohlhorst,2013]
The purpose of the data clustering is to explore the natural communities between points, textures and objects there.
These groups can be different in volume, distribution and extent of coverage.
Clustering algorithms divided to two main group, hierarchical structure and Division-based structure
hierarchical algorithms have descending or ascending function, which make cluster according to similarities from different points and based on the survey between clusters , Until they can be merged, they will be merged together .
but in other side in division approach, An initial clustering is performed on all or part of data that this cluster continues as chain
in article [Barrachina et al,2014] Based on various studies have shown that hierarchical algorithms are suitable for Data with a smaller volume and normally in data with high volume , Division-based approaches applied.
One of the most popular algorithms of this method is K-Means that used in different data group
This algorithm is one of the main data mining algorithms which are in different software packages as open codes .Various implementations have been done and different applications have been offered on the basis of this[Jain,2010]
This algorithm is very simple and fast, but two problem that arise, firstly the performance of this algorithm is largely dependent on the initial cluster centers Which arises primarily for first clusters
The second problem arises is The number of clusters that is specified with K and must be determined at the beginning of the algorithm.[Jaatun et al,2009]
This algorithm is also used in a lot of big data and Items related to cloud environments must also be considered

Various methods to improve the parallelism of the algorithm, introduced in Big Data and with their improving, also can improve specific parameters in the environment.

3- Challenges and parameters of data management in the cloud environments

The main purpose of using cloud memory, is security and reliability of data that proposed in applications due to data volume and volume of supply and demand and also Scalability
But if the data are based on definitions provided for massive data in this category, new challenges arise, as the follow:

- Flow processing: One of the issues that must be addressed in the management of massive data is that data has been sent as a sequence from different parts of data, which must be responded immediately and data management frame works must be defined dynamically that can respond dataflow immediately.

- Tolerance for Error and accessibility.

In big data , error in data management is unavoidable and the system must be such that the can tolerate faults on the other side in cloud environment system should be such that has high availability In big data , These two factors can be considered in determining management strategies together
Normally, when the data is huge, redundancy is used for accessibility
But in the case of massive data, should benefit from other strategies such creating control points for data management against hardware failures

- Scalability:

A cloud environment provides hardware resources such as memory and processing needs.
On the other hand it is necessary that communications and data aggregation methods cover data entrance and their volume without any performance drop.
These two factors can be considered in determining management strategies together
Normally, when the data is huge, redundancy is used for accessibility
But in the case of massive data, should benefit from other strategies such creating control points for data management against hardware failures
In this case, load balancing techniques on different nodes Play an important role.
Traditional techniques which is used in this field Should be developed in the required framework with using the massive data management tools that can respond system's main need as a combination.

- Effective benefit of Results obtained in parallel processing:

In big data its necessary to do division in highest level and process in parallel and after processing all results, can final result obtained.
In this part of cloudy environments, data integrity proposes that with use of multifunctional tools of data management can cover this integrity.

- Big data processing is such that parallel processing in high level is needed and high accessibility to memory and Complex calculations of computational nodes arises
Control operations and the write / read from memory in applications, done in high volume
For this reason, the operation needs to be done in a way in cloud environments that improve performance, and responding has been done in optimum time, and in Concurrent operations the ability to control access is possible.

- With using of the control level based on defined policies in cloud environments, and possibility of multilevel access of these environments, big data challenges of sharing data will be responded greatly.
But it's necessary to consider accesses in Block-level data.

- Data volume can caused problems in data locating, which can affects system performance.

Due to cloud environment's specifications, data identification must be in such way that related automatic configuration has been done based on determined policies , And as the same be done in different data centers that without any interaction with user accesses done based on labeling and keys.

4- Clustering methods in big data

In general, clustering in big data is divided into two general classes, a group of them performed on one machine and some techniques on few machines has performed.
These techniques are more popular because in Scalability fields act better and respond users optimally.

Each of these main groups has sub-group which to be described as follow:

4-1: A machine-based technique

4-1-1: Sample-based techniques

These algorithms, were the first that were used to improve the speed and scalability [Barrachina et al,2014]
These algorithms instead of clustering in whole data , consider Examples of data and after clustering extend results in all data, and because doing clustering algorithm in smaller part of data, increase the speed and have less complexity ,the most famous of them, the clustering algorithm of large applications based on random sampling, which named CLARANS shortly and can be noted.
This algorithm decrease performance time of all data and has O(n2) complexity ,This algorithm improved in most of sources[Meyer et al,2004] [Ernst et al,2005] ,to has less complexity. The other proposed algorithm in this case is BRICH.This algorithm has its special data structure and does clustering based on this structure
This algorithm has two main steps that In the first step, identify the data points and place in the tree structure of memory and then based on leaves of this tree does clustering .
CURE [Iglesias et al,2013] is another sample of such algorithms which consider a group of Scattered points as a cluster.
This algorithm proceeds hierarchically, and at first use one point as one cluster, and then with calculating distance between points of different cluster and finding minimum distance integrates different clusters [Shirkhorshidi et al, 2014]

4-1-2:Dimension reduction techniques

it is tried In these methods by reducing the dimensions , reduce the complexity of the data and improve runtime, in sample based algorithms , The emphasis is on the volume of data and there is no thought to Dimension Data [Jain,2010]

These algorithms with using of algebraic operations think to reduce the dimensions.
In this method first from Record in the original data set, the correlation matrix is calculated
Then special values with respect to the matrix to be extracted, these special values will multiplied in initial matrix .
Finally, the primary feature is a new space image ,and these properties are sorted based on them clustering features will happen[Boutsidis at al,2010]

4-2: Several machine-based techniques

According to the increasing volume of data in recent years, Strategies proposed which using computing machine to perform cannot manage Challenges related to computing power processor and required memory.

Because of this, Several machine-based technique is proposed.
In this way data divided to smaller parts and distributed between machines, and each machine individually with computing power and memory undertakes Operations on a small piece of data, these methods divided into two main groups in big data.

- Distribution of parallel non-auto-scaling

- Automatically distribution of map/reduction

In parallel clustering , Developers not only involve parallel clustering challenges, but directly faced to problems of data distribution between different nodes[Jain,2010] Which may be very complex and time-consuming, but is second group is not same and such matters managed by management workflow frame works which simplifies operations.

In general Steps in this technique is that in first step data divided and distributed between different machines ,after that each machine clusters its own related data.
The main challenges for parallelism and Clustering is To minimize data traffic and less accuracy in compare of case which serial processing is done in serial form.
This less accuracy is caused by two reasons , one is that maybe different algorithms of clustering perform in different nodes and because of this Data interference occurred

On the other side maybe change data and reflect has done incorrectly is different nodes.
Following this discussion non-automated distribution techniques, will be examined and about automatic mapping reduction algorithms will be explained in details that in next chapter combination solution to be introduced.

4-2-1: Distribution of parallel non-auto-scaling

The next stage of sample-based algorithms ,These algorithms offered however in terms of development are complex, But in terms of speed and efficiency are greatly improved One of such algorithms are DBDC [Anchalia et al,2013]
The algorithms are based on volume and distribution.
The algorithm, based on the density of points distribution occurs , This method uses density clusters
And use while there is no outlier points in data
Another example of this algorithm is ParMETIS , which acts as a multi-level and is distributed METIS algorithm ,It works as a top-down and based on the graph and from largest graph comes to smaller one.

4-2-2: Automatically distribution of map/reduction

One of the algorithms which used in the field, is PK-Means that in fact is distributed version of well-known K-Means algorithm .
The target of this algorithm is that K cluster create in data and parts of one cluster has more similarity in compare of outer points.
In first step based on determined quantities of clusters, the first cluster has created and in repeatable step , each point placed in most suitable cluster based on its distance with central point, and then central point of each cluster based on distance with other points will be updated and again second step will be repeated
PK-Means is developed of this algorithm which based on mapping/reduction framework, distributed data between several machines to speed up scalability and processing
Another algorithm in this field which suggested recently is MRDBSCAN [Mirkin,2012] which inspired from DBSCAN [He et al,2014]
This algorithm is scalable and is suggested based on map /reduce method.
First algorithms of DBSCAN couldn't manage parallelization between different nodes successfully also was limit is scalability and used architecture has not portability and parallelism, and because of this these items considered in its developed algorithm [Mirkin,2012]

5- K-means introduction

This method is basic approach for clustering which will be used in development of other approached.
The algorithm is implemented in various forms but fixed routine is respected which to be discussed in this chapter
The main purpose of these algorithm is, clustering data which the least similarity between the cluster and closest similarity inside cluster exist[Shah et al,2014]
The meaning of cluster middle distance is , distance between cluster center and dedicated points of that cluster. At first in this algorithm, K value has been considered as input and some points are chosen randomly and a primary clustering occurred and based on nearness of distance , clusters has been updated and after updating , central points of clusters with using average change and this cycle repeated again .
This process continues until no other change occurs in clusters. The best-known measure of distance calculation in cluster method is Euclidean distance criterion and the Hamming distance, which obtained respectively of the relationship of (1) and (2).
In above relation, n, shows number of dimensions and XK and YK, and point X and Y indicate K

$$d(x,y) = \sqrt{\sum_{K=1}^{n} [((XK - YK)^2])}$$

Euclidean distance relationship

$$d(x,y) = \sum_{K=1}^{n} |XK - YK|$$

Hamming distance relationship

5-1: Problems related to K-means algorithm

Final result is related to the initial cluster centers and this choose is affective in general performance, this matter will be noted in suggested and implement method.
There is no procedure for calculating the primary cluster.
In this way, it is assumed that the number of clusters is given , And this algorithm in the initial state depends to the number of clusters.

5-2: Solutions

In 2014 Lee and his colleagues offered an approach for optimizing K-mean algorithm for implementing via map-reduce function .
In this way some Changes imposed on the initial cluster centers to improve function operation distance between sample points are calculated based on the Euclidean formula and keep in one matrix named D.
A collection based on the minimum distance is initialized and C collection to be considered as Cluster's center.

Then next steps of algorithm regarding to D matrix will be done and the second minimum distance in each cluster with average distance between all points will be compare ,if it was less than average value, will be added to A collection as minimum distance and then goes to third minimum value point in D matrix and if it was more than average value , will be added to C and center of cluster and there is no need to recalculating of sample point's distance
These performance has been replete until C members become to K all number of clusters.
In same year Su and his colleagues in an article [Su et al,2014]for optimizing K-means algorithm used some ways to reduce dependence and improves function performance .
This change of performance occurs based on speed optimizing in algorithm performance.
In this method, this has been discussed after Map-reduce function that integrate results optimally .
On that year Chadha [Chadha et al,2014]and his colleagues offered an approach which based on that, dependency to K value will be lost and there is no need to input.
In this case two cluster created and the repeat based on K-means classic algorithm
And on each cluster, The new data center calculation and distance calculation will be done that no outlier point exist .
Mentioned methods are used as input for solving problems of dependency to clusters number.
As well as to improve the dependence on data centers, in a specific implementation can choose several methods for random selection or based on their parameters which can apply different methods, dynamically .
However, for greater reliability in the implementation of clustering algorithms, we can use indexing methods as combination.

6- Discussion and conclusion

In this article, the most popular clustering algorithm was evaluated, and based on done works, a method has been used which with improving primary clustering and eliminate the dependence on K, entrance operations decrease and first clusters suitable input, speed and performance will be optimized and better responding will be possible , Also with regulating the initial clustering, tolerance to error improved.

References:

V. Hristidis, S. Chen, T. Li, S. Luis, and Y. Deng, (2010),"Survey of Data Management and Analysis in Disaster Situations", the Journal of Systems and Software, Vol. 83, No. 10, PP. 1701-1717-01

K.Grolinger,(2013), "Disaster Data Management in Cloud Environments", PHD Thesis, the University of Western Ontario, PP21-27.

D. Kossmann and T. Kraska,(2010), "Data Management in the Cloud: Promises, State-of the Art, and Open Questions," Datenbank-Spektrum, Vol. 10, No. 3, PP. 121-129.

H.Dewan, R.Hansdah,(2011),"A Survey of Cloud Storage Facilities", IEEE World Congress,PP 224-231.

W. Halton,(2011), "Security Issues and Solutions in Cloud Computing", Wolf Halton.[Online], (http://wolfhalton.info/2010/06/25/security-issues-and- solutions-incloud-computing).

F.Chung, J.Dean,(2008), "BigTable: a Distributed Storage System For Structured Data",ACM.

W.Guo, C.Qiao, Y.Jin,(2010), "Demonstration of Joint Resource Scheduling in an Optical Network Integrated Computing Environment", IEEE Communications Magazine, Vol. 48, No.5, PP.7683.

W.Vogels,(2009), "Eventually Consistent", ACM, Vol 52, No.1, PP.444-0.

P. Zadrozny and R. Kodali,(2013)," Big Data Analytics using Splunk, Berkeley", CA, USA: Apress.

G.Decandia, D.Hastorun, M.Jampani,(2007), "Dynamo: Amazon highly Available Leyvalue Store", SOSP, PP.2220-05.

M. Stonebraker and et al,(2007), "The End of an Architectural Era: (it's Time for a Complete Rewrite)", 33rd International Conference on very Large Data Bases, PP.1151160-0.

F. Ohlhorst, (2013),"Big Data Analytics: Turning Big Data into Big Money", Hoboken,N.J, USA: Wiley.

A. Barrachina,A. O'Driscoll,(2014), " A big data methodology for categorising technical support requests using Hadoop and Mahout",Journal of big data,pp.11-20.

A.Jain,(2010)," Data clustering: 50 years beyond K-means", Pattern Recognition Letters,pp. 651–666.

M.G. Jaatun, G. Zhao, and C. Rong (Eds.),(2009),: CloudCom 2009, pp. 674–679.

F.Meyer, J.Chinrungrueng,(2004),"Spatiotemporal clustering of fMRI time series in the spectral Domain". Med. Image Anal. 9(1),pp. 51–68.

J.Ernst, J.Nau, Z.Bar-Joseph,(2005),"Clustering short time series gene expression data". Bioinforma.21 (suppl. 1), pp. 159–168.

F.Iglesias, F.Kastner,(2013),"Analysis of Similarity Measures in Times Series Clustering for the Discovery of Building Energy Patterns". Energies 6(2),pp. 579–597.

A.Shirkhorshidi et al,(2014)," Big Data Clustering: A Review", ICCSA 2014, Part V, LNCS 8583, pp. 707–720.

C.Boutsidis,C.Chekuri, T.Feder,R. Motwani,(2010), "Random projections for k-means clustering", NIPS, pp. 298–306.

G.Andradeand etal,(2013)," G-DBSCAN:A GPU Accelerated Algorithm for Density-based Clustering",Procedia Comput. Sci. 18,pp.369–378.

P.Anchalia, A.Koundinya,N. Srinath,(2013),"MapReduce Design of K-Means Clustering Algorithm",2013 International Conference on Information Science and Applications(ICISA), pp. 1–5.

B.Mirkin,(2012),"Clustering for data mining a data recovery approach", CRC Press.

Y.He and et al, (2014), "MR-DBSCAN: a scalable MapReduce-based DBSCAN algorithm for heavily skewed data. Front", Comput. Sci. 8(1), pp. 83–99.

N.Shah,S. Mahajan,(2014)," Distributed Document Clustering Using K-Means", International Journal of Advanced Research in Computer Science and Software Engineering, Volume 4, Issue 11,pp.24-29.

Z.Le and et al,(2014)," K-means Clustering Optimization Algorithm Based on MapReduce", International Symposium on Computers & Informatics (ISCI 2015),pp. 198-203.

S.Su and et al,(2014)," Data Clustering Using Variants of Rapid Centroid Estimation", Evolutionary Computation, IEEE Transactions, Volume: 18 , Issue: 3 ,pp. 366 – 377.

A. Chadha, S.Kumar, "An Improved K-Means Clustering Algorithm: A Step Forward for Removal of Dependency on K", 2014 International Conference on Reliability, Optimization and Information Technology, ICROIT2014.