ISSN (Online) : 2454-4159
International Journal of Advanced Research in Science Management and Technology
IJARSMT
Volume 2, Issue 2, February 2016

A Survey On Clustering Techniques For Mining Big Data

Prachi Surwade[1], Prof. Satish S. Banait [2]

ME Student, Dept. of Computer Engineering, KKWIEER, Nashik, Maharashtra, India[1]
Assistant Professor, Dept. of Computer Engineering, KKWIEER, Nashik, Maharashtra, India[2]

ABSTRACT— Clustering technique is mining process in which whole dataset is divided in to meaningful subclasses. It is unsupervised classification of data items (feature vectors (FV) observations, or patterns unsupervised) into teams (cluster). Clustering process is very useful in numerous pattern classification, grouping, exploratory pattern analysis, machine learning, document retrieval, image segmentation and decision making. "Big data" is nothing but the datasets whose size in on many facets the power of typical software package data tools to efficiently capture, store, manage and analyze. That is we tend to outline massive information in terms of being bigger than meticulous verity of thousands of gigabyte (terabyte). Big data is useful for populace and corporations but in some cases it is difficult to store and it is also time consuming. So, one of the ways to overcome this problem is to improve the clustering methods, however it suffers from high computational complexity. Data mining is the technique in which helpful information and hidden relationship among data is extracted, but the traditional data mining approaches cannot be directly used for big data due to their inherent complexity. Key objective is to introduce a simple general overview of data clustering categorizations for big data. And also explain and summarizing some of the related work for it. This paper presents a theoretical overview of some of current clustering techniques used for analyzing big data.

KEYWORDS- *Data mining, Big data, Clustering techniques, Big data analytics.*

I. INTRODUCTION

Clustering techniques is the commonly used techniques. It is mainly used to analysis of data. Clustering process combining group of items and grouping can be done by creating the most similar item of one group (cluster) and most non similar items in the other group. This process is similar to classification. Clustering is an unsupervised way of learning or it is an unsupervised classification of data items, (feature vectors (FV), patterns, or observations) into clusters (groups). Clustering is found the clusters (or grouping) of data in a set of unlabeled data. In artificial intelligence (AI) data clustering technique is major assignment. Human being can easily identify the cluster in low dimension with less number of records, but in case of computer it is extremely hard to instruct the computer to find such a relationship. When the dimensions of the data increases human has difficulties in finding the interesting patterns of the data. To find an interesting pattern in exploratory data analysis or to extract the information from the data is the objective of exploratory data analysis.

We are living in the 21st century, the digital age. Every day, people store large information and it is representing as data for further analysis and management. The amount of data in our world has been growing regularly. Company takes

ISSN (Online) : 2454-4159
International Journal of Advanced Research in Science Management and Technology
IJARSMT
Volume 2, Issue 2, February 2016

a million of bytes of data related to their consumers, dealers and their related operation, and trillions of n/w sensors are used to establish (set) in the real world in storage spaces or devices like automobiles and mobile phone, for creating, and sensing and communicating data it needs smart phones or social networking sites or other devises that will used to maintain data exponential expansion. "Big data" can be defined as a large datasets whose size is so (too) large for the database software tools cannot easily capable to store data, capture data and handle data. We do not define big data in terms of being larger than a certain number of thousands of gigabytes (terabytes) [2]. Recently technology advancement over time period, the size of datasets increases we called as big data. Therefore, big data analytics can be good to impact business change and improve results, by applying advanced analytic techniques on big data, and discovering hidden insights and helpful information.

At the starting of new era information has grown-up speedily in sizes as well as in varieties also. This data collected is of huge amount and it has some degree of difficulty when it comes to collecting and analyzing data as big data. Data mining method used to extracts useful information and unseen correlation between data. Huge amount of data stored in datasets are quickly growing due to rapid technological growth (progress). Therefore some applications require the storage space and repossession of complex multimedia items that may be represented by high dimensional FV. Very difficult task is to obtain more valuable information concealed in some databases therefore to remove these difficulty Clustering analysis used. It is best techniques which are applied on large data sets for analyzing purpose.

Clustering is one of the most important issues in data mining and machine learning [3]. Clustering is a task of discovering homogenous groups of the studied objects. Many researchers have a significant interest in developing clustering algorithms. The most important issue in clustering is that we do not have prior knowledge about the given data. Moreover, the (input) parameters choice like number nearest neighbors, Kn amount of clusters and some other factor in algorithms create the clustering is challenging process. One of the very effective ways of dealing with these data is to categorize or assemble that data into a set of classes. Now a day's clustering methods emerge as another influential meta-learning tool for correctly analyze the big volume of data created by some new applications. The Big data can also be define as the datasets which having big dimensions or they are in large variety as well as in large velocity so it is very hard to hold that datasets by applying conventional techniques and tools. Just because of some fast expansion of information, we require solutions that efficiently handles and extract knowledge from these datasets. Therefore analysis of clustering techniques with their some different available classes with big datasets provides an effective and useful conclusion.

II. BIG DATA

Big data includes or big data nothing but all information to be stored in only one database and also providing those data security .big data also used in hadoop concept.

What is Big Data?

Big data is large no of data to be stored into only one warehouse and providing security.

In big data is collections of datasets are used.

2.1 What Comes Under Big Data?

Big data involves the data produced by different devices and applications. Given below are some of the fields that come under the umbrella of Big Data[3].

ISSN (Online) : 2454-4159
International Journal of Advanced Research in Science Management and Technology
Volume 2, Issue 2, February 2016

1) Black Box Data:

In black box data usually complicated electronic devices whose internal mechanism is hide from user.

For e.g. Component of helicopter, airplanes, and jets, etc. in that capture their voice and performance.

2) Social Media Data:

All sites of data or all forms of data collected in social media .social data are mostly in Facebook and Twitter.

Stock Exchange Data: In stock exchange data keep the information about product sell and product buy.

3) Power Grid Data: Power grid data holds information consumed by a particular node with respect to a base station. Or providers and consumers are connected by transmission and distribution lines or data operated by one or more centers.

4) Transport Data: Transport data nothing but Transfer all data files containing all actual data copied into target location or nothing but export and import data.

5) Search Engine Data: A data having big information about the particular (sites) or at the time of knowing the information that time retrieved the data associate with user information.

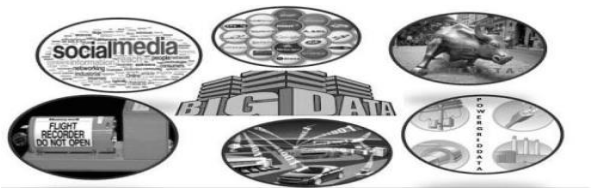

Fig 1 Big Data Overview

2.2 Characteristics of Big Data

Big data characteristics listed below:

i. Volume – The volume is related to the size of data. At present data is in Petabytes and in near future it will be of zettabytes.

ii. Variety – The data is not coming from single source it includes semi structured data like web pages, log files etc., raw, and structured and unstructured data.

iii. Variability – The variability considers inconsistencies in data flow.

iv. Value – The value is importance of data used by the user. The user queries against certain data stored, obtains result, rank them and can store for future work.

v. Velocity – The velocity is related to the speed and all data are coming from different resources.

vi. Complexity – The data is coming from various resources in huge amount thus it is difficult to link or correlate multiple data.

2.3 Big Data Analytics Challenges

Below are the types of big data challenges [4]:

1. Volume of data is large and also varies so challenge is how to deal with it.
2. Analysis of all data is required or not.
3. All data needs to be stored or not.
4. To analyze which data points are important and how to find them.

ISSN (Online) : 2454-4159
International Journal of Advanced Research in Science Management and Technology
Volume 2, Issue 2, February 2016

5. How data can be used in the best way

3. Big data clustering technique

Similar properties grouping objects in clustering process. Any cluster should be split in two properties;

1. Low intra class similarity.
2. High intra class similarity.

Clustering is an unsupervised learning technique and it is learn by observation. There are no predefined class label exits for the data point. Cluster analysis used for so many types of applications like image processing and market processing etc. Clustering technique is used for determine the intrinsic grouping in a set of the unlabeled data. And all similar data or similarity between data object can be measured with imposed distance value.

3.1 Techniques of Big data:

Generally, Big Data clustering techniques are 3 types :

A. Single machine clustering technique.

B. Multiple machine clustering technique.

C. Hybrid Clustering Technique

This technique used for or used to empower a clustering algorithm and in this to work on bigger datasets through improve their speed and scalability. Following Figure 2 shows the process for develop in clustering to deals with big data.

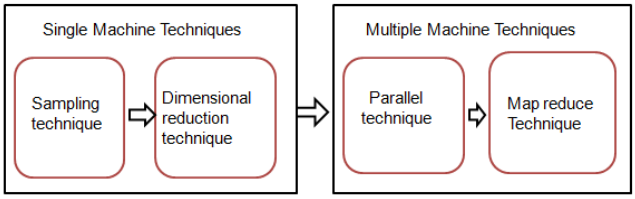

Fig 2. Process for develop in clustering to deals with big data

Big Data, single-machine clustering algorithm or techniques used in single and it can use resources of one single machine. Multiple machines clustering algorithms or techniques used on multiple machines and it can use resources of multiple machines. These two methods of big data include some different techniques illustrated in Figure3

A) Single Machine clustering algorithm

1) Sample Based Clustering

In this algorithm we can used for improve the speed and scalability of dataset and there target deal with its search space is called sample based algorithm because it perform clustering algorithm on sample of dataset instead of performing clustering on whole dataset and then generalise that sample dataset as whole dataset. It will depends on their speed and scalability. There are many different sample based clustering methods [5] given below.

Partition Based Clustering (PBC)

In PBC techniques the dataset is partitioned into various clusters. By partitioning each data element would have the cluster index. In this approach user must have to pre-define the (kn) some number of clusters with some criteria parameters and on the basis of this parameters solution will be evaluated. The most popular algorithms of partitioning cluster are Partitioning Around Mediods (PAM), Clustering Large Applications method (CLARA), K-Mean method. The well known distance method for this category is Euclidean distance method for K-Means.

ISSN (Online) : 2454-4159
International Journal of Advanced Research in Science Management and Technology
Volume 2, Issue 2, February 2016

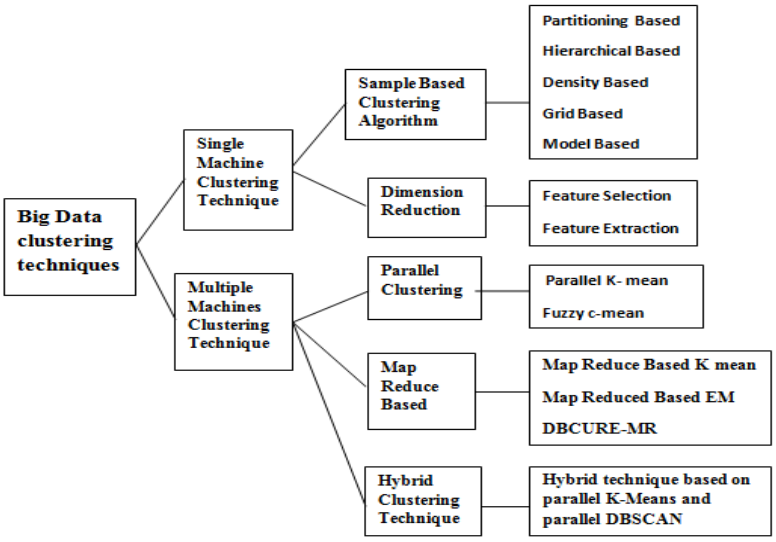

Fig 3.Big Data clustering Technique

Hierarchical Based Clustering (HBC)

In HBC, clustering tree is build or we can say clusters hierarchy is build which represent the result of cluster analysis. In this the prior number of clusters is not to be mentioned. HBC technique categorize in two methods as agglomerative and divisive method. They use a bottom up and top down approaches respectively. In top down approach each cluster is split into different number of clusters and in bottom up approach clustering is starts with one cluster and it combines (merges) two or more clusters. The most popular algorithm of this category is BIRCH, CURE, ROCK and Chameleon is some algorithms.

Density-Based Clustering (DBC)

The purpose of DBS method is to determine arbitrary shaped clusters. This method separates the high density regions and low density region of data objects of the cluster. This algorithm is also used for determining connected graph as it checks the distance between core point with other data points and checks whether distance is not less than the defined radius. The density of each data point is calculated by detecting the Kn number of data objects (items) in its neighborhood. The clusters are said to be dense if it would has more than minimum points; the minimum points are the number of data objects (points) which should be present in the cluster. The most common technique of this category is DBSCAN which deals with noisy data as well.

Grid-Based Clustering (GBC)

The GBC method creates a grid configuration means it divides a data spaces in definite number of cell. In the gird configuration different operation for clustering is performed. These methods always have high-speed processing time. This method is independent on available data points (items) and depends on gridcells of grid configuration. Main advantage of GBC method is the significant reduction in complexity. The most common algorithm of this category is: GRIDCLUS, STING, CLICK and Wave Cluster.

Model-Based Clustering (MBC):

MBC method is used to optimize a fit between predefined algebraic (mathematical) models and predefine data. Data can be created by using a fusion of underlying probability distribution is the hypothesis of MBC method. Based

ISSN (Online) : 2454-4159
International Journal of Advanced Research in Science Management and Technology
IJARSMT
Volume 2, Issue 2, February 2016

on statistics it determines number of clusters and it takes noises in account. MBC method uses a statistical approach (SA) and neural network approach (NNA). For determine the clusters SA uses a probability measures where probability measures is used to characterize every derived clusters. NNA uses a group of joined I/O units and each connected I/O unit associated with weight. MCLUST (Model based clustering) method is most popular algorithm of this category. Another best method is EM (Expectation Minimization) method uses a mixture of density model.

2) Dimension Reduction:

In dimension reduction data size measured in two dimensions and the number of variables. These two dimensions take very high values, which could cause a problem during the exploration and analysis of these data. For this, it is essential to implement data processing tools and make a pretreatment to the dataset before applying clustering algorithms for a better understanding knowledge available in this data.

The purpose of dimensionality reduction is to select or extract optimal subset of relevant features for a criteria already fixed. The selection of this subset of features can eliminate irrelevant and redundant information according to the criterion used. This selection or extraction makes it possible to reduce the size of the sample space and makes it all more representative of the problem . For large sets of data, dimension reduction is usually performed before applying the classification algorithm to avoid the disadvantages of high dimensionality .Most of two types of dimensionality reduction technique:

2.1) Feature selection:

It aims to select an optimal subset of variables from a set of original variables, according to a certain performance criteria. The main objective of this selection is to reduce the number of required actions. A work made in 2014 [6] proposes a classification algorithm for Big Data based on feature selection.

2.2) Feature extraction:

It aims to select features in a transformed space - in a projection space, the extraction methods use all the information to compress and produce a vector of smaller dimension[7].

B. Multiple-machine clustering:

1) Parallel clustering: The processing of large amounts of data imposes a computing to achieve results in reasonable time. In this section, we examine some parallel algorithms

and distributed clustering used to treat Big Data; the parallel classification divides the data partitions that will be distributed on different machines. This makes an individual classification to speed up the calculation and increases scalability. Parallel clustering algorithms applied to any of applications using clustering algorithms for efficient computing. Parallel algorithms add difficulty for distribution. It is worth full because of the major improvement in scaling and speed of clustering algorithm. It involves not just parallel clustering challenges but also detail in data distribution process between different machine available in network [8]. Parallel clustering algorithm and distributed clustering used to treat with big data ,in which parallel classification divides data partitions that will be distributed on different machine. This makes an individual classification to speed up the calculation and increases the scalability. P k-mean is distributed version of K-mean algorithm. P k-mean has almost linear speed up, also linear size up and also has good scale up[9].

2) Map Reduce based clustering:

ISSN (Online) : 2454-4159
International Journal of Advanced Research in Science Management and Technology
IJARSMT
Volume 2, Issue 2, February 2016

Map Reduce is a task partitioning mechanism (with large volumes of data) for a distributed execution on a large number of servers[10]. Principle is to decompose a task (the map part) into smaller tasks. The tasks are then dispatched to different servers, and the results are collected and consolidated (the reduce part).The function of this framework is shown in Fig. 4.

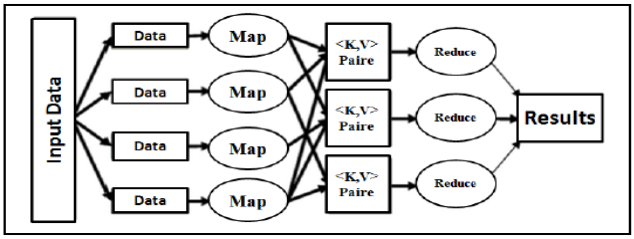

Fig. 4 Map Reduced Framework

In step Map the input data is analyzed, cut into sub problems and delegated to other nodes (which can do the same recursively). This will be processed later using the Map function which has a pair (key, value) that associates a set of new pairs (key, value). Then comes the stage Reduce, where the lowest nodes reach their results back to the parent node that had asked them. It calculates a partial result using the Reduce function (reduction) involving all the corresponding values for the same key to a unique pair (key, value). Then he goes back information in turn. In this Category Map Reduced Based K mean, Map Reduced Based EM[11], DBCURE-MR [12] algorithm is proposed:

C. Hybrid Clustering Technique:

Hybrid clustering technique based on parallel K-Means or parallel DBSCAN that combines the benefits of both parallel K-Means and parallel DBSCAN algorithms. The benefit of parallel K-Means is that it is not complex and forms clusters in less execution time while the advantage of parallel DBSCAN is that it forms the cluster of arbitrary shape. Figure 4.1 shows that the main procedure of proposed method that works in three stages;

1. Stage 1 (PARTATION): The first stage is the data partition stage in which data is partitioned into different regions or parts to minimize the boundary points.

2. Stage 2 (MAPPER) : In the second stage mapper function is performed on each node in which local clusters are formed.

3. Stage 3 (REDUCER) : The third stage is the reducer stage in which mean of each cluster is calculated and it is returned as the cluster id of each cluster.

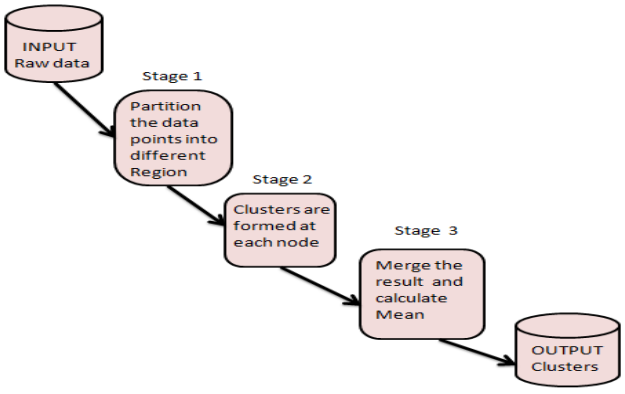

Fig 4. Execution of Hybrid Big data clustering technique

ISSN (Online) : 2454-4159
International Journal of Advanced Research in Science Management and Technology
IJARSMT
Volume 2, Issue 2, February 2016

III. LITERATURE SURVEY

T. Zhang, R. Ramakrishna [13], suggested a BIRCH is an HBC algorithm. This algorithm is categorized into agglomerative approach. For Big databases this algorithm is useful. It is especially suitable for very large databases. This algorithm has been planned for minimizing number of I/O operations. BIRCH algorithm is dynamically and incrementally clusters the incoming data items (objects) and then they try to create a greatest quality of clusters with in some limited assets (as time constraints and available memory). In this algorithm memory data-structure uses called as clustering feature tree (CF-tree). First clustering method is BIRCH. It is firstly developed for handle noise. Disadvantage of this method is it only handles numeric data.

S. Guha , R. Rastogiand K. Shim [14], present CURE's agglomerative HBC method. In this algorithm random data samples are 1st partition. After that every created partition is then partly cluster. Next final pass for creation of final clusters to preclustred an data from each partitions by eliminating an outliers. In each step of this algorithm, closest pair of each clusters medoid (centroids) of two clusters is combining (merged). This method only combines representative objects (centroid) and uses single linkage method for selecting more than one centroid from each cluster. Disadvantage of this method is combined inter-connectivity of points in two different clusters information is not considered. To overcome this problem Chameleon algorithm [15] is implemented .

E.-H. Han, V. Kumar and G. Karypis [15], suggested "CHAMELEON" as HBC agglomerative method which can discover dynamic modeling. This algorithm work on two phases : 1st divides (partition) data items into a sub clusters then using graph partitioning method continually merging sub-clusters and then to get a final clusters. This algorithm is mainly used to find clusters densities their diverse shapes and sizes in 2D (two dimensional) space. To obtain the arbitrary shape clusters and arbitrary densities of clusters Chameleon method uses an dynamic model. The algorithm is applicable for the applications whose data volume is large. Disadvantage of CHAMELEON only is that it is well-known for low dimensional data space.

K. Shim, S. Guha and R. Rastogi [16], discussed ROCK HBC technique. This agglomerative HBC technique uses a idea of links [8]. It is suitable for managing big volume data sets. In ROCK technique similarity of clustering depends on the points of diverse clusters which are nearest in ordinary can measures clusters similarity. This algorithm is employes a link not a distance for merging clusters. Also they discussed the next version of the ROCK algorithm is QROCK algorithm used for clustering a categorical data. QROCK is quicker than QROCK [17].

J. Han and R. T. Ng [18], proposed CLARANS algorithm. CLARANS technique supports to randomized search. It is not used any predefined configuration. This algorithm not more affect on incrementing dimensionality of database. This technique does not require distance function. It supports point as well as polygonal objects. Also researcher discussed that existing algorithm is as PAM and CLARA is less efficient than CLARANS.

K-means clustering algorithm is classifying or grouping items into k groups where K is the number of pre-chosen groups). The grouping is done by minimizing the sum of squared distances or Euclidean distances between items and the corresponding centroid [19].

ISSN (Online) : 2454-4159
International Journal of Advanced Research in Science Management and Technology
IJARSMT
Volume 2, Issue 2, February 2016

K-Mean has Advantages is that, if large numbers of variables are present then K-Means may be computationally faster than hierarchical clustering. If the clusters are globular then K-Means may produce tighter clusters than hierarchical clustering. Similarly K-Mean has some disadvantages is that -

1) It is Difficult to comparing quality of the produced clustered.
2) Fixed number of clusters can make it difficult to predict what K should be.
3) It does not work well with non-globular clusters

DBSCAN(Density-Based Spatial Clustering of Applications with Noise) was proposed to adopt density-reachability and density connectivity for handling the arbitrarily shaped clusters and noise [20].But DBSCAN is very sensitive to the parameter Eps (unit distance or radius) and MinPts (threshold density), because before doing cluster exploration, the user is expected to estimate Eps and MinPts.

DENCLUE (Density-based Clustering) is a distribution-based algorithm [21], which performs well on clustering large datasets with high noise. Also, it is significantly faster than existing density based algorithm, but DENCLUE needs a large number of parameters.

Map Reduce is a most popular model use for processing large set of the data.it offers numbers of benefits to handle large data sets such as scalability, flexibility and fault tolerance.Map reduce framework is widely used in processing and managing large data sets. It is also used in such applications like document clustering, access log analysis, and generating search indexes.

Parallel K-means has been studied by [Dhillon], [Xu], [Stoffle][22] previously for very large database. In K-mean algorithm te speed up and scale up variations respect to document (vectors).K-mean algorithm the dimension of each of the documents has been studied [Dhillon]. In parallel K-means algorithm used for the distributed memory multiprocessors was implemented for SPMD model.For testing purpose K-means algorithm also used or probabilistic variants of K-means as well as for very large categorical datasets K-models[Huang] Variant of k-means could also be parallelized.

IV. SUMMARY AND CONCLUSION

This paper introduces big data and provides background of various clustering techniques used to analyze big data. The objective of this paper was to survey the most important clustering algorithms and determine which of them can be used for clustering large datasets. In this study, we reviewed some papers about the clustering algorithms and the corresponding parallel clustering algorithms. Parallelism in the clustering algorithm has been used for both efficient clustering strategy and efficient distance computation. Based on literature survey, there are various big data clustering techniques which are used to analyse big datasets but these techniques are not efficient as some of them are related to particular task and they do not provide the global solutions, some of them are fast but they had to compromise with the quality of clusters and vice versa. Therefore to overcome this problem hybrid clustering algorithm is proposed and designed that helps in analysing big data in an efficient manner. In future, we can be applied all big data clustering techniques for a particular application area by addressing issues involved and can be applied for data sets with categorical attributes.

ISSN (Online) : 2454-4159
International Journal of Advanced Research in Science Management and Technology
IJARSMT
Volume 2, Issue 2, February 2016

REFERENCES

[1] Btissam Zerhari, Ayoub Ait Lahcen, Salma Mouline1, "Big Data Clustering: Algorithms and Challenges", CONFERENCE PAPER • MAY 2015F. Chung, Spectral Graph Theory. Providence, RI, USA: American Mathematical Society, 1997.

[2] S. Suthaharan, M. Alzahrani, ``Labelled data collection for anomaly detection in wireless sensor networks," in Proc. 6th Int. Conf. Intell. Sensors, Sensor Netw. Inform. Process. (ISSNIP), Dec. 2010, pp. 269_274.

[3] A.K. Jain, M.N. Murty, and P.J. Flynn, "Data Clustering: A Review," ACM Computing Surveys, vol. 31, no. 3, pp. 264-323, Sept. 1999.

[4] A. Katal, M. Wazid and R.H. Goudar, "Big data: Issues, challenges, tools and goodpractices," Contemporary Computing (IC3), 2013 Sixth International Conference on,IEEE, 2013.

[5] R. Xu and D. Wunsch, "Survey of clustering algorithms.," IEEE transactions on neural networks / a publication of the IEEE Neural Networks Council, vol. 16, no. 3, pp. 645-78, May. 2005.

[6] 12 F. Bu, Z. Chen, Q. Zhang, and X. Wang, "Incomplete Big Data Clustering Algorithm Using Feature Selection and Partial Distance," InDigital Home (ICDH), 5th International Conference on. IEEE, p. 263-266, 2014.

[7] B. J. Kim, "A Classifier for Big Data," In Convergence and Hybrid Information Technology. Springer Berlin Heidelberg, p. 505-512, 2012.

[8] Manasi N. Joshi," Parallel K - Means Algorithm on Distributed Memory Multiprocessors" Spring 2003 Computer Science Department University of Minnesota, Twin Cities

[9] K. Stoffel and A. Belkoniene, "Parallel k/h-means clustering for large data sets," In Euro-Par'99 Parallel Processing. Springer Berlin Heidelberg, p. 1451-1454, 1999

[10] Donald Miner and Adam Shook" MapReduce Design Patterns", Printed in the United States of America.

[11] Y. Zhao, Y. Chen, Z. Liang, S. Yuan, and Y. Li, "Big Data Processing with Probabilistic Latent Semantic Analysis on MapReduce, International Conference on Cyber-Enabled Distributed Computing and Knowledge Discovery, 162 – 166, 2014.

[12] K. Younghoon, S. Kyuseok, K. Min-Soeng, L. June Sup, "DBCUREMR: An efficient density-based clustering algorithm for large data using MapReduce," Information Systems, vol. 42, p. 15-35, 2014.

[13] T. Zhang, R. Ramakrishnan, and M. Livny, ``BIRCH: An efficient data clustering method for very large databases," in Proc. ACM SIGMOD Rec., Jun. 1996, vol. 25, no. 2, pp. 103_114.

[14] S. Guha, R. Rastogi, and K. Shim, ``Cure: An ef_cient clustering algorithm for large databases," in Proc. ACMSIGMOD Rec., Jun. 1998, vol. 27, no. 2.

[15] G. Karypis, E.-H. Han, and V. Kumar, ``Chameleon: Hierarchical clustering using dynamic modelling," IEEE Comput., vol. 32, no. 8, pp. 68_75, Aug. 1999.

[16] S. Guha, R. Rastogi, and K. Shim, ``Rock: A robust clustering algorithm for categorical attributes," Inform. Syst., vol. 25, no. 5, pp. 345_366, 2000.

[17] M. Dutta, A. Kakoti Mahanta and A.K. Pujari, QROCK: A quick version of the ROCK algorithm for clustering of categorical data, Pattern Recognition Letters, 26 (2005), 2364-2373.

[18] R. T. Ng and J. Han, ``CLARANS: A method for clustering objects for spatial data mining," IEEE Trans. Knowl. Data Eng. (TKDE), vol. 14, no. 5, pp. 1003_1016, Sep./Oct. 2002.

[19] ALSABTI K., RANKA S., SINGH V., " An Efficient k-means Clustering Algorithm, Proc. First Workshop High Performance Data Mining, 1998.

[20]]Ester M., Kriegel H.-P., Sander J., Xu X.: "A Density- Based Algorithm for Discovering Clusters in Large Spatial Databases with Noise", Proc. 2cnd Int. Conf. On Knowledge Discovery and Data Mining, Portland, Oregon, 1996, AAAI Press, 1996.

[21] A. Hinneburg and D. A. Keim, ''Optimal grid-clustering: Towards breaking the curse of dimensionality in high-dimensional clustering,'' in Proc. 25th Int. Conf. Very Large Data Bases (VLDB), 1999, pp. 506–517.

[22] Dhillon , Xu , Stoffle]and A. Belkonience. "Parallel k-Means clustering for large datasets". Proceedings of EuroPar -1999.