Big Data Clustering: A Comparative Study On Various Clustering Algorithms

G Ashok Kumar

Abstract: Analysts classify big data as volume, velocity, and variety. Big data analysis explores intelligence from extremely wide variety of dynamic and complex data. Data cleaning is an essential step in big data analytics for easy prediction / decision making / clustering using data organizing tools. Clustering performs grouping of similar data from a population data set so that the data points in the same group show high degree of similarity between them than to the data points of other groups. Big data clustering help researchers to perform dimensionality reduction in complex problems, designing spam filters, identifying fraudulent or criminal behavior, performing Document analysis, classifying network traffic and helping Marketing /Sales analysis. The paper makes analysis of prominent big data clustering techniques in classifying data points belonging to different level of complexities.

Index Terms: Big Data, Clustering, Data Cleaning, Dimensionality Reduction, Analysis, Volume, Velocity, Variety and Dynamic.

1. INTRODUCTION

After a period of managing data accumulation challenges, these days the issue is reformed into how to progress with these enormous measures of data. Researchers and scientists trust that today a standout amongst the most imperative topic in computer science is Big Data. For example: Social networks such as, Facebook and Twitter have many billions of clients and they create gigabytes of data every moment, retail locations ceaselessly gather their clients' information, You Tube has one billion exclusive clients which are delivering hundred hours of cinematic video each an hour and its substance ID service look over four hundred years of video consistently. Big data refers to huge information as a mixture of relevant and irrelevant data for a particular application. Data mining identifies interesting patterns from big data by performing "knowledge discovery in databases (KDD)". Figure 1 portrays the fundamental architecture of Big data. Huge Data are tied in with turning unstructured, precious, defective, complex data into usable data. Be that as it may, it ends up hard to keep up large volume of data and information every day from various assets and administrations which were not accessible to human space only a couple of decades prior. To manage this torrential slide of data, it is important to utilize amazing assets for information discovery. Clustering is one of them that is considered as a strategy in which data are separated into groups such that objects in each group offer more likeness than with different articles in different collections. Data clustering is an outstanding system in different zones of software engineering and related areas. Despite the fact that data mining can be considered as the primary cause of clustering, however it is limitlessly utilized in different fields of learning, for example, machine learning, bio informatics, networking, energy engineering, pattern recognition and along these lines a great deal of research works has been done here. From the earliest starting point researchers were managing clustering methods so as to deal with their complexity and computational expense and thus increment adaptability and speed.

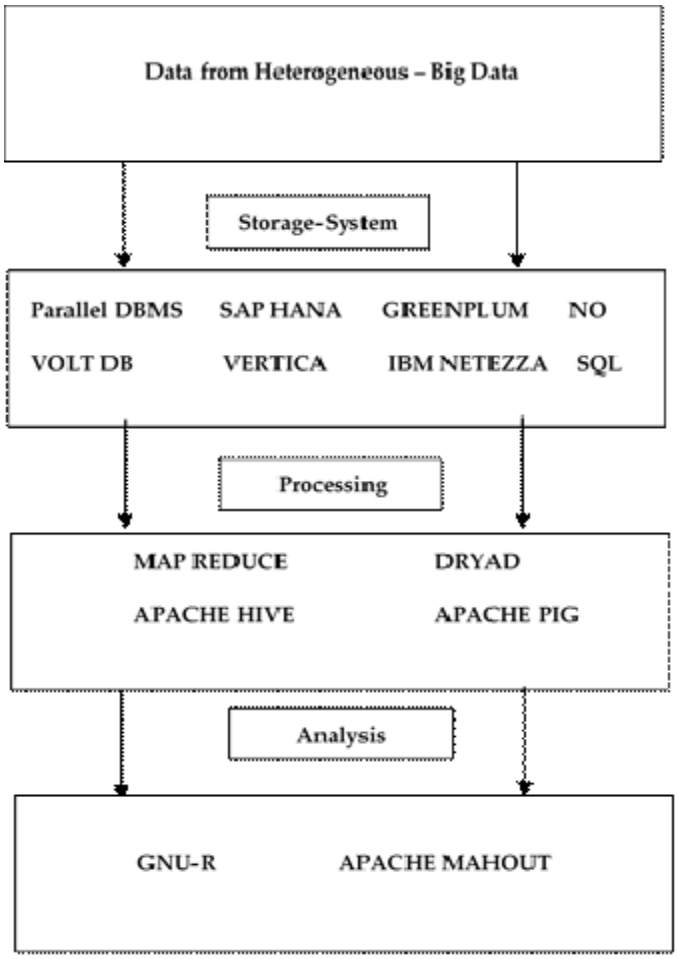

Figure 1. *Architecture of Big Data*

Emersion of Big data added more difficulties to this subject which asks more research for clustering methods enhancement. This paper presents a sensible investigation of various Clustering strategies to development information which encourages researchers to pick which grouping strategy is ideal to group the data dependent on the prerequisites.

- *Ashok Kumar G is currently working as an Assistant professor in Karpagam College of Engineering, PH-+919894168777. E-mail: rgashokkumar@gmail.com*

2 CHALLENGES IN BIG DATA

The procedure of information available on the web progresses exponentially. All in all, we determine that the information can be well-ordered into three fundamental sorts: structured, unstructured and semi-structured. The vast majority of created data are unstructured and conventional database tools can't deal with this sort of data.

2.1 Volume

Now a days, masses of data to be handled are exponentially expanding. This depicts the way that our expanded utilization of new innovations (cell phones, social media...) urges to create an ever-increasing number of information in our day by day events both specific and specialized; the organizations are challenging a burst of stored information. In reality, this dimension keeps on evolving at rapid. It is evaluated that the capacity of information kept in the world doubles at even intervals. It has warehoused a bigger number of information in 2010 than other previous years.

2.2 Velocity

Velocity in the aspect of big data specifies the speed of incoming data for analysis. Velocity in the aspect of big data specifies the speed of incoming data for analysis. Information regarding Velocity is essential for big data analytics for data formation, processing, filing data catching, trading and retrieving. So, the regular data collection, examination and utilization are essential for effective data analytics.

2.3 Variety

Variety in the aspect of big data measures the diversity of incoming information. Without a doubt, it can utilize the information contained in sites, web journals, messages, trades on interpersonal organizations (Facebook, LinkedIn, Twitter, ...), the biometrics, pictures, video, sound, logs, geolocation, and so forth. Their birthplaces are various: mining picture, web, content mining, and so on. We have to consolidate a few foundations to reach important inferences. The collection of Big Data illuminates the distress of applying the information or data from customary data warehousing architecture. In realism, the unexpected test of Big Data is to make assorted information (geolocation, climate, vehicle traffic, logistics) and partner them to extricate valuable data and therefore enhance the different areas exploiting this colossal amount of information wide and scattered. The greatest significant features of Big Data are:

- Heterogeneity refers to the variety in sources and data representation types. It also provides the information regarding data interconnectivity, interrelation, represented. The data representation can be of structured, semi-structured, and even entirely unstructured and hence contextualization heterogeneous data streams is essential for improving the performance of classification.
- Autonomous, Autonomous refers to the level of one's freedom of working on large data set without seeking any permission / authority. It spread the controls so that every information or data source can work spontaneously without being originated on any centralized control.
- Complexity, Complexity in big data measures the complexity of incoming data in terms of data structures, sources, types and processing such as sequential / parallel processing. The complexity level is directly proportional to the volume and types of data. Data complication increments with the expansion in volume and the typical action strategies, with the organization of common database tools are not ever again acceptable to meet the fundamentals catch, storage and further investigation.
- Evolving, the growth of composite information additionally speaks to a basic component. Enormous information is changing rapidly.

To deal with the developing requests of information, we should develop or grow the boundary and execution of devices and techniques. Massive information or Data involves new resolutions for increase the limit and gripping action to exploit practically of the information or data without select the new properties or resources. Unquestionably, with the exponential growth of information or data, outdated information mining techniques have been not able to address critical issues as far as information preparing. So as to exploit this huge volume of information, active dealing out model with a workable computational outflow of this tremendous, dynamic, complex, and heterogeneous information or data is required.

3 DEFINITION OF CLUSTERING

Clustering is an unsupervised learning errand where one tries to recognize a limited arrangement of classifications named groups to portray the data. Clustering is likewise characterized as "A gathering of the equivalent or comparable components accumulated or happening firmly together". Clustering divides the population data set to different clusters depending on the homogeneity of different classes. Different metrics are used by different clustering algorithm to perform clustering of data. For example, figure 2 shows some real time clusters. Any incoming object will be classified in to different groups depending on the extracted features of the given object to frame clusters. The grouping process utilizes inter and intra class similarities of different objects to make the final decision to group clusters. Large differences in features of incoming objects with existing clusters result in introducing new clusters. A better Clustering technique will create top quality groups with high intra-class comparability - Similar to each other inside a similar group low between class likeness - Different to the objects in different groups.

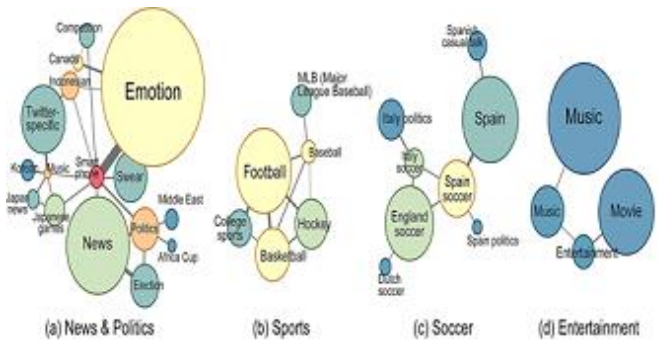

Figure 2. *Examples of cluster*

Clustering techniques utilizes similarity information among different objects to perform effective classification. The effectiveness of clustering techniques depends on how well it

classifies a new object to a new / existing class / unseen pattern.

4 VARIOUS TYPES OF CLUSTERING ALGORITHMS

In general, the big data clustering frameworks can be described into two important classes: single-machine grouping procedures and multi machine grouping methodologies. Various machine grouping systems has pulled in more consideration since they are increasingly adaptable in versatility and offer quicker reply time to the clients. As it is exhibited in Figure 3 single-machine and multi machine grouping strategies includes different methods:

- Single-machine clustering
  - Datamining based clustering
  - Dimension reduction techniques
- Multiple-machine clustering
  - Parallel clustering
  - MapReduce based clustering

4.1 Single Machine Clustering

4.1.1 Data Mining Clustering Algorithms

Machine learning algorithms are classified into supervised / unsupervised clustering techniques in which a supervised algorithm uses its trained experiences while unsupervised methods utilize the visible similarity / differences of current objects. The goal of unsupervised methods needs to ensure that the objects assembled in a similar group are comparable and predictable as indicated by explicit constraints. It is hard to put on information or data mining grouping measures in Big Data due to the new problems. As the amount of data / information increases, the computational complexity, handling and examination costs of clustering algorithms increases. Further, these complex situations with high volume of information increase the burden of clustering algorithms to produce effective clusters in limited time. These clustering strategies can be divided into: partitioning based, model-based, density based, hierarchical based and grid-based model.

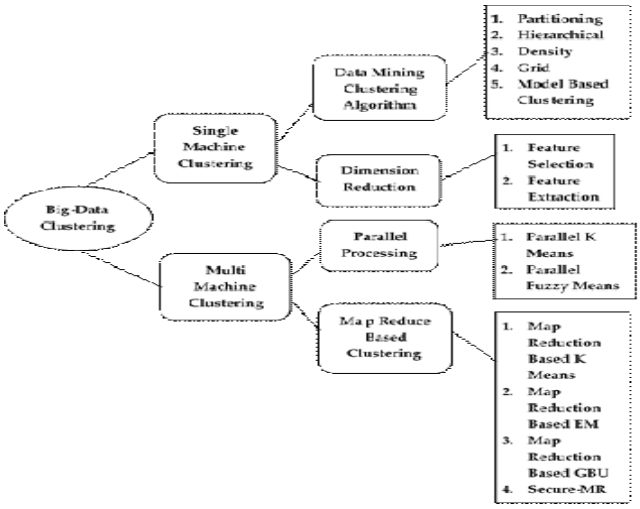

Figure 3. *Different Clustering Algorithms*

4.1.2 Partitioning Based Clustering Algorithms

All things are considered from the start as a lone collection. The things are sheltered into number of assignments by iteratively discovery the areas between the segments. Partitioning technique behaviors one - level partitioning on informational collection, first it makes starting arrangement of k segment, where parameter k is the quantity of segments to build. It at that point utilizes an iterative movement strategy that endeavors to enhance the partitioning by moving items starting with one group then onto the next group. Usually partitioning technique incorporates two prominent algorithms called, k - means and k - medoids. In many real time situations, the information regarding the optimum number of clusters will be unavailable and hence k-means clustering algorithm requires additional pre-processing step to determine the number of clusters. When these inputs are online with no prediction of future data, k means clustering are not practical. There are many subdividing techniques, for example, CLARA, CLARANS, K-implies, FCM, PAM, and k-medoids K-modes.

4.1.3 Hierarchical Based Clustering Algorithms

This strategy partitions information into various dimensions that take after a level of importance. This grouping gives an unmistakable data visualization. There are two ways to deal with perform Hierarchical grouping procedures, 1. top-base and 2. Base top. In top-base system, from the outset one article is picked and dynamically consolidates the neighbor objects reliant on the distance as least to maximum level. The procedure is consistent until the point that an ideal cluster is surrounded. The base top procedure oversees set of items as single group and segments the cluster into further groups until needed number of groups are molded. BIRCH, CURE, SNN, ROCK, Chameleon, GRIDCLUST, Wards, Echidna, and CACTUS are some of hierarchical grouping methods. The various hierarchical technique has a noteworthy disadvantage, which is identified with the way that once a phase is finished, it can't be fixed.

4.1.4 Density Based Clustering Algorithms

The clustering method dependent on density can discover groups in a discretionary way, where the groups are described as solid regions disconnected by low compactness zones. Clustering methods reliant on density are not proper for tremendous informational collections. Information objects are requested into focus points, outskirt focuses, and noise focuses. All the inside points are related together subject to the densities to shape cluster. Subjective formed clusters are confined by various grouping calculations, for instance, OPTICS, DBSCAN, GDBSCAN, DBCLASD, DENCLU and SUBCLU. There are two methodologies in density-based clustering. The primary methodology pins density to a trained information point and the special techniques join OPTICS and DBSCAN. The subsequent technique pins compactness to a point in the trademark space and it incorporates the method DENCLUE. The main demerit of density-based strategy is an absence of interpretability.

4.1.5 Grid Based Clustering Algorithms

Grid based approaches segment the information or data set into numeral of cells to outline a framework structure. It adjusts to the three stages: partitioning, erasing unwanted information and consolidating important features. The phenomenal preferred point of view of grid-based clustering is the

enormous diminishing in multifaceted nature. Contrast with all Clustering methods Grid-based algorithms are quick handling methods. Unchanging Grid-based calculations are not acceptable to outline needed groups. To beat these issue Adaptive Grid-based designs, for MAFIA, instance, and AMR Arbitrary formed groups are moulded by the grid cells. A few precedents are: CLIQUE, BANG, OptiGrid, ENCLUS, PROCLUS, ORCLUS, STIRR, FC and STING and WaveCluster.

4.1.6 Model Based Clustering Algorithms

Set of information points are related together subject to various methods like numerical procedures, conceptual systems, and solid clustering strategies. There are two measures for model-based techniques, one is numerical strategy, and another is neural framework approach. The principle weakness of this methodology is moderate processing time if there should be an occurrence of vast data sets. Algorithms, for example, COBWEB, EM, CLASSIT, SOM, and SLINK are outstanding Model based clustering techniques.

4.2 Dimension Reduction

In spite of the fact that the intricacy and speed of grouping methods is identified with the quantity of instances in the dataset, however at the other pointer dimensionality of the dataset is another persuasive viewpoint. Truth be told the more dimensions data have, the more is complication nature and it implies the more drawn out execution time. The quantities of factors and models take high appearances, which could cause an issue in the midst of the check-up and valuation of these information. For this, it is critical to understand the information or data making tools and make a pre-treatment to the dataset before applying grouping techniques. The determination or removal makes it likely to decline the size of the model space and makes everything increasingly illustrative of the issue.

4.2.1 Feature Selection

Usually the data for analysis contains many unique features and incorporating all features drags the classification algorithm to fail. Hence it is essential to select optimum number of features to provide faster results. The right off the bat, the feature selection technique is expected to diminish the degree of the dataset. By then, an equal k-means calculation is associated with the information subsets picked in the underlying advance.

4.2.2 Feature Extraction

The feature selection process identifies the most prominent features used for performing classification while feature extraction processes the given data to capture most relevant information in the form of reduced dimensionality representation. Many component extraction systems rely upon PCA, LS-SVM, and also., the exploratory results on the Big Data exhibits that the proposed clustering calculation reliant on feature extraction permits taking care of large classification issues.

4.3 Multi Machine Clustering

These days the development of data size is way a lot quicker than recall and processor progressions, subsequently one machine with a solitary processor and a recollection can't deal with terabytes and petabytes of information and it underlines the need methods that can be kept running on various machines. This procedure permits to breakdown the tremendous amount of data into lesser pieces which can be stacked on various machines and afterward utilizes processing power of these machineries to take care of the huge issue.

4.3.1 Parallel Clustering

The handling of a great deal of information forces a parallel processing to achieve results in sensible time. In the primary stage, information will be isolated into segments and they disperse over machines. A short time later, each machine performs grouping independently on the allotted partition of data. Two principle contests for parallel and distributed grouping are limiting information traffic and its lower precision in examination with its sequential equivalent. In parallel clustering, developers are included with parallel clustering difficulties as well as with subtleties in information distribution process between various machines accessible in the system too, which makes it exceptionally confounded and tedious. This element permits huge parallelism and less demanding and quicker adaptability of the parallel framework.

4.3.2 Map reduce based clustering

Though parallel grouping algorithms enhanced the versatility and speed of clustering calculations still the multifaceted nature of managing memory and processor conveyance was a calm essential challenge. MapReduce is a distributed computing model that works in Map and Reduce stages. Map stage of the algorithm breaks down the given set of data to individual tuples while reduce stage combines the output tuples of map stage to smaller set of tuples. Typical MapReduce algorithm framework contains three operations:

- Speed up: implies the proportion of running time while the information set stays consistent and the quantity of machines in the framework is expanded
- Scale up: measures if x time bigger framework can perform x time bigger job with a similar run time
- Size up: keeping the quantity of machines unaltered, running time develops straightly with the size of data.

5 DISCUSSION

The clustering algorithms which fulfill the vast majority of the criteria are looked at and broke down in Table 1 and 2 based on the 3V's. This clustering depends on their steadiness, execution and adaptability execution. From this investigation it very well may be reasoned that there is no productive order calculation as indicated by all assessment criteria, yet we take note of that EM and FCM demonstrated better execution as far as quality than alternate methods. Each one of those strategies are appearing for a lot of information determined time is detonating. Along these lines, so to solution for this issue we should exploit from an effective programming dialect or propelled specialized gear. The calculation BIRCH, DENCLUE and OptiGrid are increasingly adjusted to a lot of information however they experience the ill effects of low clustering quality.

TABLE 1 COMPARISON BASED ON VOLUME

Comparative Methods	Volume		
	Data Set Classification	Dimensionality (High)	Avoidance of Outliers
BANG [36]	Large	Large	Yes
BIRCH [38]	Large	No	No
CACTUS [43]	Small	NO	No
Chameleon [7]	Large	Yes	No
CLARA [20]	Large	No	No
CLARANS [15]	Large	No	No
CLASSIT [17]	Small	No	No
CLIQUE [23]	Large	No	Yes
COBWEB [18]	Small	No	No
CURE [25]	Large	Yes	Yes
DBCLASD [34]	Large	No	Yes
DBSCAN [9]	Large	No	No
DENCLUE [35]	Large	Yes	Yes
ECHIDNA [16]	Large	No	No
EM [10]	Large	Yes	No
ENCLUS [30]	Large	No	Yes
FC [8]	Large	Yes	Yes
FCM [12]	Large	No	No
GDBSCAN [27]	Large	No	No
GRID- CLUST	Small	No	No
K- medoid [37]	small	Yes	Yes
K-means [8]	Large	No	No
k-modes [28]	Large	Yes	No
MAFIA [11]	Large	No	Yes
OPTICS [33]	Large	No	Yes
OptiGrid [2]	Large	Yes	Yes
ORCLUS [1]	Large	Yes	Yes
PAM [44]	Small	No	No
PROCLUS [39]	Large	Yes	Yes
ROCK [26]	Large	No	No
SLINK [22]	Large	No	No
SNN [5]	Small	No	No
SOM's [21]	Small	Yes	No
STING [24]	Large	No	Yes
STIRR [8]	Large	No	No
SUBCLU [32]	Large	Yes	Yes
Wards [31]	Small	No	No

TABLE 2 COMPARISONS BASED ON VARIETY AND VELOCITY

Comparative Methods	Variety		Velocity
	Data Set Classification	Shape of the Cluster	Computational Complexity
BANG [36]	Numerical	Arbitrary	O(n)
BIRCH [38]	Numerical	Non convex	O(n)
CACTUS [43]	Categorical	Hyper rectangular	O(c N)
Chameleon [7]	All types data	Arbitrary	O(n2)
CLARA [20]	Numerical	Non convex	O(k(40+k)2+k(n-k))
CLARANS [15]	Numerical	Non convex	O(kn2)
CLASSIT [17]	Numerical	Non convex	O(n2)
CLIQUE [23]	Numerical	Arbitrary	O(C k + m k)
COBWEB [18]	Numerical	Non convex	O(n2)
CURE [25]	Numerical &Categorical	Arbitrary	O(n2logn)
DBCLASD [34]	Numerical	Arbitrary	O(3n2)
DBSCAN [9]	Numerical	Arbitrary	O(n log n)for spatial data
DENCLUE [35]	Numerical	Arbitrary	O(log |D|)
ECHIDNA [16]	Multi- variate	Non convex	O(N*B(1+logB m))
EM [10]	Spatial	Non convex	O(knp)
ENCLUS [30]	Numerical	Arbitrary	O(ND+ m D)
FC [8]	Numerical	Arbitrary	O(n)
FCM [12]	Numerical	Non convex	O(n)
GDBSCAN [27]	Numerical	Arbitrary	no
GRID- CLUST	Numerical	Arbitrary	O(n)
K- medoid [37]	Categorical	Non convex	O(n2dt)
K-means [8]	Numerical	Non convex	O(n k d)
k-modes [28]	Categorical	Non convex	O(n)
MAFIA [11]	Numerical	Arbitrary	O(cp + p N)
OPTICS [33]	Numerical	Arbitrary	O(n log n)
OptiGrid [2]	Spatial	Arbitrary	O(n d) to O(nd- logn)
ORCLUS [1]	Spatial	Arbitrary	O(d3)
PAM [44]	Numerical	Non convex	O(k(n-k)2)
PROCLUS [39]	Spatial	Arbitrary	O(n)
ROCK [26]	Numerical &Categorical	Arbitrary	O(n2+nmm- ma+n2logn)
SLINK [22]	Numerical	Arbitrary	O(n2)
SNN [5]	Categorical	Arbitrary	O(n2)
SOM's [21]	Multi variant	Non convex	O(n2m)
STING [24]	Spatial	Arbitrary	O(k)
STIRR [8]	Categorical	Arbitrary	O(n)
SUBCLU [32]	Numerical	Arbitrary	no
Wards [31]	Numerical	Arbitrary	no
Wave Cluster [14]	Numerical	Arbitrary	O(n)

6 CONCLUSION

This paper describes different clustering methods and distinguishes the data clustering algorithms used to oversee the extensive arrangements of data. For the most part, so as to oversee huge volume of information, the clustering algorithms need to be enhanced by decreasing their time and space complexities. The survey revealed that CLIQUE and

BIRCH and ORCLUS algorithms provide better performance in data clustering of big data analytics in the presence of large number of outliers. The present survey suggests that effective clusters can be formed by using CURE and ROCK strategies on ordered information. For forming effective discretionary clusters spatial information strategies such as OPTIGRID, STING, PROCLUS and ORCLUS can be incorporated in clustering algorithms. The main aim of this examination is the determination of good clustering algorithm as clustering system is most famous strategy for the grouping of objects and utilized in numerous fields like Biology, Libraries, Insurance, and Marketing and so on.

REFERENCES

[1] Abzetdin Adamov. Distributed file system as a basis of data-intensive computing, in: 2012 6th International Conference on Application of Information and Communication Technologies (AICT), pp. 1–3 (October).

[2] Breunig M, Ankerst M, Kriegel HP, Sander J. Optics: Ordering points to identify the clustering structure. Proceedings of the ACM SIGMOD International Conference on Management of Data. 1999 Jun; 28(2):49–60.

[3] C. YADAV, S. WANG, M. KUMAR, “Algorithm and approaches to handle large Data-A Survey,” International Journal of computer science and network, vol 2, issue 3, 2013.

[4] C.K. Reddy, C.C. Aggarwal, Data Classification: Algorithms and Applications. CRC Press, 2014.

[5] Chatterjee S, Sheikholeslami G, Zhang A. Wave cluster: A multi resolution clustering approach for very large spatial da-tabases. Proceedings Int Conf Very Large Data Bases (VLDB); 1998. p. 428–39.

[6] D. WUNSCH and R. XU, “Survey of clustering algorithms,” Neural Networks, IEEE Transactions, vol. 16, no 3, p. 645-678, 2005.

[7] Dr. Amit Ganatra, Prof. Neha Soni1, Comparative study of several Clustering Algorithms, International Journal of Advanced Computer Research, Volume-2 Number-4 Issue-6 December-2012.

[8] Ehrlich R, Bezdek JC, Full W. FCM: The Fuzzy C-Means Clus-tering algorithm. Computers and Geosciences. 1984; 10(2-3):191–203.

[9] Erchart M, Schikuta E. The BANG – Clustering system: Grid–based data analysis. Lecture Notes in Computer Science. 1997; 1280:513–24.

[10] Ester M, Xu X, Sander J Krieger HP. A distribution-based clustering algorithm for mining in large spatial databases. Pro-ceedings 14th IEEE International Conference on Data Engi-neering (ICDE); Orlando, FL. 1998 Feb 23-27. p. 324.

[11] Eui-Hong (Sam) Han, George Karypis, Vipin Kumar, Chameleon: Hierarchical Clustering Using Dynamic Modeling, Computer, v.32 n.8, p.68-75, August 1999 [doi>10.1109/2.781637.

[12] G. Q. Wu, X. Wu, X. Zhu, and W. Ding, “Data mining with Big Data,” Knowledge and Data Engineering, IEEE Transactions on, vol. 26, no 1, p. 97-107, 2014.

[13] Grobelnik M, Brank J, Mladenic D. A survey of ontology eval-uation techniques. Proceedings Conf Data Mining and Data Warehouses; 2005. p. 166–9.

[14] Han EH, Karypis G, Kumar V. Chameleon: Hierarchical clus-tering using dynamic modeling. IEEE Computer. 1999 Aug; 32(8): 68–75.

[15] Han J, Ng RT. Efficient and effective clustering methods for spatial data mining. Proceedings Int Conf Very Large Data Bases (VLDB); 1994. p. 144–55.

[16] Han J. CLARANS, Ng RT: A method for clustering objects for spatial data mining. IEEE Transactions on Knowledge Data Engineering (TKDE). 2002 Sep/Oct; 14(5):1003–16.

[17] http://quantumcomputers.com.

[18] http://www.whitehouse.gov/sites/default/files/microsites/ostp/big-data-fact-sheet-final-1.pdf.

[19] J Macqueen. Some methods for classification and analysis of multivariate observations. Proceedings 5th Berkeley Sympo-sium on Mathematical Statistics Probability; Berkeley, CA, USA. 1967. p. 281–97.

[20] Jun CH , Park HS. A simple and fast algorithm for K-me-doids clustering. Expert Systems Applications. 2009 Mar; 36(2.2):3336–41.

[21] Karmasphere Studio and Analyst, 2012. <http://www.karmasphere.com/>.

[22] Keim DA, Hinneburg A. An efficient approach to clustering in large multimedia databases with noise. Proceedings ACM SIGKDD Conf Knowl Discovery Ad Data Mining (KDD); 1998. p. 58–65.

[23] Kriegel HP, Ester M, Sander J, Xu X. A density-based algo-rithm for discovering clusters in large spatial databases with noise. Proceedings ACM SIGKDD Conf Knowl Discovery Ad Data Mining (KDD); 1996. pp. 226–31.

[24] Leckie C, Mahmood AN, Udaya P. An efficient clustering scheme to exploit hierarchical data in network traffic analy-sis. IEEE Transactions on Knowledge. Data Engineering. 2008 Jun; 20(6):752–67.

[25] M.B.Vaidya, Yaminee S. Patil, A Technical Survey on Cluster Analysis in Data Mining, International Journal of Emerging Technology and Advanced Engineering Website: www.ijetae.com (ISSN 2250 - 2459, Volume 2, Issue 9, September 2012).

[26] M.Renuka Devi, M.Vijayalakshmi, A Survey of Different Issue of Different clustering Algorithms Used in Large Datasets, International Journal of Advanced Research in Computer Science and Software Engineering, Volume 2, Issue 3, March 2012.

[27] Markus M, Mihael Ankerst, Breunig, Hans-Peter Kriegel, Jörg Sander, OPTICS: ordering points to identify the clustering structure, Proceedings of the 1999 ACM SIGMOD international conference on Management of data, p.49-60, May 31-June 03, 1999, Philadelphia, Pennsylvania, United States.

[28] P Berkhin. Survey of clustering data mining techniques in grouping multidimensional data. Springer. 2006; 25–71.

[29] P. Ahlawat MANN and P. Batra NAGPA, “Survey of Density Based Clustering Algorithms,” International journal of Computer Science and its Applications, vol. 1, no 1, p. 313-317,2011.

[30] Pentaho Business Analytics, 2012. <http://www.pentaho.com/explore/pentaho-business analytics/>.

[31] Philip Bernstein, Divyakant Agrawal, Elisa Bertino, Susan Davidson, Umeshwas Dayal, Michael Franklin, Johannes Gehrke, Laura Haas, H.V. Jagadish, Jiawei Han Alon Halevy, Alexandros Labrinidis, Sam Madden, Yannis Papakon stantinou, Jignesh Patel, Raghu Ramakrishnan, Kenneth Ross, Shahabi Cyrus, Dan Suciu, Shiv Vaithyanathan, Jennifer Widom, Challenges and Opportunities with Big Data, CYBER CENTER TECHNICAL REPORTS, Purdue University, 2011.

[32] Rajeev Rastogi, Sudipto Guha, Kyuseok Shim, CURE: an efficient clustering algorithm for large databases, Proceedings of the 1998 ACM SIGMOD international conference on Management of data, p.73-84, June 01-04, 1998, Seattle, Washington, United States.

[33] Ramakrishna R, Zhang T, Livny M. BIRCH: An efficient data clustering method for very large databases. Proceedings of the ACM SIGMOD International Conference on Management of Data. 1996 Jun; 25(2):103–14.
[34] Rastogi R, Guha S, Shim K. Cure: An efficient clustering algo-rithm for large data bases. Proceedings of the ACM SICMOID international Conference on Management of Data. 1998 Jun; 27(2):73–84.
[35] Rastogi R, Guha S, Shim K. Rock: A robust clustering algo-rithm for categorical attributes. 15th International Conference on Data Engineering; 1999. p. 512–21.
[36] Rousseau PJ, Kaufman L. Finding groups in data: An intro-duction to cluster analysis. USA, Johns and Sons Wiley; 2008.
[37] S. Aghabozorgi, A. S. Shirkhorshidi , T. Y. Wah, and T. Herawan, “Big Data Clustering: A Review,” In Computational Science and Its Applications–ICCSA 2014. Springer International Publishing, p. 707-720. 2014.
[38] Sheetal Sisodia, Deepti Sisodia, Lokesh Singh, Khushboo saxena, Clustering Techniques: A Brief Survey of Different Clustering Algorithms, International Journal of Latest Trends in Engineering and Technology (IJLTET). Vol. 1 Issue 3 September 2012.
[39] Storm, 2012. <http://storm-project.net/>.
[40] Tari Z, Fahad A, Alshatri N, Alamri A. A survey of clustering algorithms for Big Data: Taxonomy and empirical analysis. IEEE Transactions on Emerging Topics in Computing. 2014 Sep; 2(3):267–79.
[41] Wang S, Yadav C, Kumar M. Algorithms and approaches to handle large data sets - A survey. International Journal of Computer Science and Network. 2013; 2(3):1–5.
[42] Wunsch D , Xu R. Survey of clustering algorithms. IEEE Trans-actions on Neural Networks. 2005 May; 16(3):645–78.
[43] Xu Xiaofei, He Zengyou , Deng Shengchun, Squeezer: an efficient algorithm for clustering categorical data, Journal of Computer Science and Technology, v.17 n.5, p.611-624, May 2002.
[44] Z Huang. A fast clustering algorithm to cluster very large cate-gorical data sets in data mining. Proceedings SIGMOD Work-shop Res Issues Data Mining Knowl Discovery; 1997. p. 1–8.
[45] Zhai C, Aggarwal C. A survey of text clustering algorithms. Mining Text Data. New York, NY, USA. Springer-Verlag: 2012. p. 77–128.