A Study of Clustering Approaches and Validation Measures for Big Data Mining

Kamlesh Kumar Pandey[1], Diwakar Shukla[2]

[1]Research Scholar, Dept. of CSA, Dr. Hari Singh Gour Vishwavidyalaya, Sagar (M.P.) India.

[2]Prof & HOD, Dept. of CSA, Dr. Hari Singh Gour Vishwavidyalaya, Sagar (M.P.) India.

ABSTRACT

Present time natures of data are totally changed to big data with respect to volume, variety, and velocity. Clustering is one of the unsupervised approaches of data mining and data mining is one of the approaches for big data analysis and known as big data mining. Clustering is a very helpful technique under big data mining because it discovers distribution of patterns, hidden relations, self-noise and outlines management, class label predication and interesting correlations in large data sets and high dimensional data. Every traditional clustering algorithm works under specific criteria and these criteria define the cluster and validation measure validates this cluster as the requirements. From a theoretically, practically and the existing research perspective, this paper study seven clustering taxonomies such as partition, hierarchical, density, grid, model, fuzzy and graph based clustering taxonomy and their validation measures for the big data mining.

Keywords: Big Data, Big Data Mining, Clustering Taxonomy, Clustering Validation Measures, Cluster Validation Taxonomy

I INTRODUCTION

Nowadays, large scale data are generated from a different type of sources such as health, social network, government, cloud computing, e-marketing, internet of the thighs, financial, sensor network and so on. IDC predicts data volume reaches 44 Zetta-bytes (44 billion terabytes) per day till 2020, which is ten times double by 2003. In general, the big data refer to the large datasets that are collected from heterogeneous sources and these sources are continually growing. Due to the changing the nature data to big data some government establishes a big data department. In March 2012, the Obama Administration launched the Big Data Research and Development Initiative and July 2012, the Japan government established the Big Data development under the national technological strategy. The United Nations issued a report entitled "Big Data for Development: Opportunities and Challenges" which describes outlining the Big Data challenges and their dialogue (Oussous et al., 2018). In Big Data perspective, traditional data related techniques such as data management, process management, analysis technique have touched a bottleneck and cannot finish the data processing in fixed time with accuracy, gives the slow responsiveness, and problems with scalability because big data needs to high capacity for data storages, low-value density, complex dynamic relation between various data types, high processing speed, high scalability, availability and reliability (Weichen et al., 2016).

Clustering is one of the techniques for data analysis which is finding a similar relation or pattern for unlabeled objects. In big data mining, various applications such as social network analysis, customer segmentation, scientific data analysis, bioinformatics, and target marketing used to cluster analysis. The traditional clustering algorithms are not suitable under large-scale data because it takes high computation time. Consequently, computational efficiency is the biggest and most important challenge of large-scale data and need to how to improve the clustering algorithm. At the present time parallel and distributed computation are solutions to these types of the problem (Zhao et al., 2019). A good clustering algorithm in a big data environment need to cluster is more accurate and reliable. The clustering accuracy, reliability and algorithm efficiency is measured by cluster validation. This paper is mainly focused on the clustering algorithm and validation approach in the big data mining environment through five sections.

II BACKGROUND

(a) Big Data

Laney (2011) described big data through three dimensions such as volume, velocity, and variety.This dimension defines a common framework of big data.

The volume describes the size of the data, variety describes the heterogeneous sources and types of data and velocity describes the speed of data generation, analysis, and processed. Social Media is the best example of these 3V's because social media is generating high volume data in the form of high variety in the form of high velocity. Gartner et al., 2012, summarized these dimensions of big data as "Big data have high volume, high velocity, and high variety information assets that demand cost-effective, innovative forms of information processing for enhanced insight and decision making." After that various, research origination gives some support dimension for the big data framework. IBM describes veracity as a fourth dimension, which is related to unreliability and uncertainty latent in data origin for improving data accuracy and quality with the trustworthiness of data. After that SAS added variability dimensions to big data, which represents the variation in data flow rates. In addition, the variation of data flow rates is related to increasing the variety and velocity. Oracle introduced six supportable dimensions as value, which hold to hidden value or attribute during the big data mining. The last dimension is related to the visualization of all dimensions according to the user needs. Visualization is visualized the data mine and analysis results according to user expectation, such as the graph, table, or another format (Gandomi et al., 2015, Lee et al., 2017, Sivarajah et al., 2017). These all dimensions are known as 7 V's of big data. Volume, Variety, and Velocity are known as basic dimensions of big data creation and Veracity, Variability, Value, and Visualization are known as supportable dimensions of big data accession. Figure 1 summarized all dimensions of the big data.

Fig.1 7V's of Big Data

(b) Big Data Mining

The big data mining approach is totally differs to the traditional data mining approach because traditional data mining algorithm based on centralized databases, but the big data mining algorithm based on distributed or multiple sources, traditional data mining algorithm can't be able to handle huge scale data sets, high dimensional, heterogeneous data format and source and scalability. In Big Data perspective, the data mining technique must be deal high volume, high variety and high velocity with scalability.

When big data mining algorithm applied in heterogeneous sources, the method of mining is divided into four groups as the heterogeneous source with pattern analysis, classification, clustering, and fusion (Wang et al., 2018). The purpose of big data mining is not only mining the interesting knowledge's and summarizing the data, but it is also mine the consistent patterns, systematic and complex relationships among the data through classification, clustering, association rule learning, regression and other data mining techniques. Big data mining used machine learning and statistical methods for evolution, extended to utilize traditional data mining techniques (Siddiqa et al., 2016).

Big data mining does not support to the relational model because the relational model handles only structured data, but the big data needs to handle structured, unstructured and semi-structured data with the high scale and distributed natures. At the present time, NoSQL databases are more popular for data storages because it stores the heterogeneous data format. In this database, the data are stored in the distributed file and graph format. In the distributed file format model, the data are stored in the form of a key/value pair and the graph-based model has data organized in the form of a vertex/edge pair. The Graph data model is useful for evaluating the various problems such as distances computing, finding relationships, community detection, determining connectivity and so on (Weichen et al., 2016). The distributed and parallel architecture is very helpful for achieving performance and accuracy reduces the computation time and scalability for the requirements of clustering in big data mining. The MapReduce programming model is one of the approaches for distributed and parallel computing in big data mining (Chong et al., 2015, Sardar et al., 2018).

III CLUSTERING TAXONOMY

Clustering is one of the approaches for analysis and discovering the complex relation, pattern, and data in the form of underlying groups for the unlabeled data objects. The data objects of each group share the same similarity and the other group data objects totally different from another group (Zhao et al., 2019). Data clustering is the most important technique and produces high-quality analytical results for data reduction. Data clustering also increases the efficiency and accuracy of data mining under unclassified problems (Chen et al., 2018). In Big Data perspective, the clustering algorithm must be deal high volume, high variety and high velocity with scalability.

Similarity and dissimilarity (distance) are two basic functions for cluster creation. Every clustering taxonomy is used to these functions for cluster construction on the bases of own cluster creation behaviours and natures.In nowadays various Similarity and dissimilarity measure are available for clustering under the Minkowski, L(1), L(2), Inner product, Shannon's entropy, Combination, Intersection and Fidelity family (Cha et al.,2007, Kocher et al., 2017, Manning et al., 2008).

The design of the clustering algorithm under the big data mining is fulfilled the volume, velocity, and variety related criteria. Fahad et al., 2014 and Pandove et al., 2015 describes Volume related criteria such as cluster is must be dealt huge size, high dimensional and noisy of the dataset, Variety related criteria such as cluster is must be recognized as dataset categorization and cluster shape, and Velocity related criteria define the complexity, scalability, and performance of the clustering algorithm during the execution of real dataset.

In general, Clustering algorithm divided into seven groups such as partition, hierarchical, density, grid, model, fuzzy and graph based clustering taxonomy based on their cluster creation, working process, behaviors and cluster nature. The basic concept of the theses clustering algorithm is described under section **a** to **g** (Chen et al., 2018, Fahad et al., 2014, Gan et al., 2007, Pandove et al., 2015, Shirkhorshidi et al., 2014)

(a) Partitioning based method

This clustering method creates a cluster on the bases of the center data point, then reason it's known as centroid-based clustering. This method partitions the dataset using a K number of user define the cluster and randomly assigned data object in each K cluster after that it finds the center point and assigns the objects to the nearest center of the cluster. K-Mean, K-Medoids, K-parameter, PAM, CLARA, and CLARANS are the most popular clustering algorithm under this clustering approach.

(b) Hierarchical based Clustering

This technique is also known as Connectivity-based clustering because the cluster is created by using tree (hierarchy of clusters) concept. Agglomerative and Divisive are two basic methods for cluster creation under this clustering taxonomy. The agglomerative method starts from the individual data cluster because each data object has own cluster at the beginning and after that, each cluster pair of data objects is moved up to the hierarchy form until the needed cluster is not found. Divisive starts from the top cluster because all data objects belong into one cluster at the beginning and after that, the cluster is divided into different cluster pairs move down the hierarchy until the needed cluster is not found. Here cluster is shown as dendrogram format. BIRCH, CURE, ROCK, Chameleon, ECHIDNA, WARDS, and SNN are the most popular clustering algorithm under this clustering approach.

(c) Density-based method
This clustering method is creating a cluster on the bases of data object density, connectivity and borderline. These clustering techniques provide the barriers against the noisy and determine the clusters to an arbitrary shape. To find out density it used to Mean-shift concept, firstly calculate the mean of current data point and figure out them and this iteration will be continued until the desired cluster is not founded. DBSCAN, OPTICS, DENCLUE, and GDBSCAN are the most popular clustering algorithm under this clustering approach.

(d) Model-basedclustering
This clustering method creates a cluster on the bases of some existing model such as mathematically model; statically model, probability model and other distribution model, and finding the best data object fit into the model. The model is depending on the existing model, then that reason it is known as distribution-based clustering. Here each model needs to minimize according to their distribution, but the use of multiple parameters it takes high time complexity. COBWEB, SLINK, SOM, ART, and EM are the most popular clustering algorithm under this clustering approach.

(e) Grid-based clustering
This clustering method used for utilizing the space for data sets in multidimensional data. Here each original data space is separated into grids or calls structure for defining the size of the cluster. Here data space is divided into many rectangular cells by using the hierarchical structure for parallel processing for fast processing time, and the data is organized within the different cell levels. This clustering algorithm generally used for statistical techniques. STING, CLIQUE, Wave Cluster, OptiGrid, MAFIA, ENCLUS, PROCLUS, ORCLUS, and STIRR arethe most popular clustering algorithm under this clustering approach.

(f) Fuzzy based clustering
This type of clustering is based on fuzzy or soft computing or hard computing for real data clustering then reason it's called soft computing based clustering. In the hard clustering, data object must belong to only one cluster, but the soft clustering data object belongs to multiple clusters on the bases of the membership function. Fuzzy clustering provides more insight and knowledge about the data objects to all clusters. FCM, FCS, and MM are the most popular clustering algorithm under this clustering approach.

(g) Graph based clustering
This clustering algorithm is realized on the graph, where the node refers to a data point and the edge is referred to as the relationship betweenall data points. During this cluster,the analysis graph must be in the form of a minimum spanning tree. This clustering algorithm is based on traditional and spectral graph theory. CLICK and MST is the most popular clustering algorithm under this clustering approach.

IV CLUSTER VALIDATION APPROACH

Clustering validation measures evaluate the goodness of clustering results and its validity for the success of the clustering algorithm. The clustering validation measures are categorized into External clustering validation and Internal clustering validation groups (Aggarwal et al., 2014). These groups are measures which clustering algorithm is suitable for volume, variety and velocity criteria.

(a) External clustering validation
External clustering validation measures evaluate "purity" of the clusters respect to given class labels. External validation measures helpful for finding the exact number of the cluster in the advance and given the framework for choosing an optimal clustering algorithm on the particular dataset (Liu et al., 2010). The overall this approach test the points of the data set are randomly organized as pre-specified structured or not and this analysis is done by using the Null Hypothesis (Halkidi et al., 2002). Table 1 shows some popular external clustering validation measures with their equation definition (Aggarwal et al., 2014, Arbelaitz et al., 2013 Halkidi et al., 2002, Liu et al., 2010).

Entropy and purity measures are given the purity and accuracy of the class labels with respect to the cluster. F-measure is given the precision and recall value together. F-measure is useful for the information retrieval community. The Mutual Information (MI) measures depended to the random variable, this measure compares how much information is depended to the random variable and compared to another random variable.

The Variation of Information (VI) measures the quantity of information that is lost or grown in changing the form of the class set to the cluster. The Rand statistic, Jaccard coefficient, Fowlkes & Mallows index, and Hubert's statistics I and II evaluate the clustering quality of the pair of data point by using the agreements and/or disagreements in different partitions. The Minkowski score measures the difference between the clusters, which is obtained by the clustering algorithm and also given the disagreements of the data point in different partitions. The classification error is given a total misclassification rate in each class to a different cluster. It is very helpful for minimizing the total misclassification rate. The Van Dongen criterion is related to evaluating the graph clustering measures and it is given the measures of majority objects in each class and each cluster.

Table 1 shows external clustering validation measures where some mathematically formulation defines as Data set as D with n objects, assume that there is a partition $A = \{A1, \bullet\bullet\bullet, AK\}$ of D.

$Ai\ j = ni\ j/n,$
$Ai = ni\bullet/n,$
$Aj = n\bullet\ j/n.$

(b) Internal clustering validation

Internal clustering validation measures evaluate the goodness of a clustering structure with respect to trustiness on information in the data inside of the cluster without external information. Internal validation measures are useful for finding the best clustering algorithms and the optimal number of the cluster without any other information (Liu et al., 2010). Overall, this approach evaluates the clustering algorithms and their resultsby using the quantities and features to the used dataset (Halkidi et al., 2002). Table 2 shows some popular internal clustering validation measures with their equation definition (Aggarwal et al., 2014, Arbelaitz et al., 2013 Halkidi et al., 2002, Liu et al., 2010).

Table 2 shows Internal clustering validation measures where some mathematically formulation defines as D use as data set, n use as number of objects in D, c use as center of D, Ci use as the ith cluster,ni use as number of objects in Ci, ci use as center of Ci, A use as attributes number of D,k use as number of nearest neighbors, d(x, y) use as distance between x and y, NC use as number of clusters, qj use as number of Ci's jth object's nearest neighbors which are not in cluster Ci;

Compactness and separation are two basic measures for internal validation. Compactness is measured, how closely related the data objects in the cluster area by lower variance indicates and numerous measures on the base of distance, such as maximum, minimum or average for pairwise and center-based distance. Separation is measured, how distinct cluster is from other clusters for using compare cluster density, centers, and minimum distances. Such internal validate index as Davies–Bouldin (DB), Xie-Beni index (XB), and Silhouette index (S) consider both evaluation criteria compactness and separation and such as Root-mean-square standard deviation (RMSSTD), R-squared (RS), and Modified Hubert statistic (considers describe only one internal validation aspect.

The root-mean-square standard deviation (RMSSTD) validates the homogeneity of the shaped clusters by the square root pooling sample attributes. R-squared (RS) validates the degree of difference between the clusters of the sum of squares. Here, the sum of squares between clusters to validates the total sum of squares of the whole data set. The Modified Hubert statistic validates the difference between clusters by counting the disagreements of data point pairs in the partitions. The Calinski–Harabasz index (CH) validates the cluster validity by average between-and within-cluster on the basis of the sum of squares. Index I (I) validates the separation by the using maximum distance between cluster centers and given the compactness by the sum of distances between data points and their cluster center. Dunn's index (D) is given a minimum pairwise distance between different clusters data point for intercluster and the maximum diameter among different clusters data point for intracluster compactness. The Silhouette index (S) validates the clustering performance by using the pairwise difference of the cluster distances within and between the clusters. The Davies–Bouldin (DB) validity index is useful for cluster similarity obtained by averaging all the cluster similarities. The smaller index is indicating to the better clustering result and the max index is indicating the clusters are most distinct from. The Xie-Beni index (XB) validate the inter-cluster separation by the minimum square of the distance between cluster centers and also validate the inter compactness by the mean square of the distance between all data points and its cluster center. SD index (SD) validates average scattering and the total separation of clusters. It validates compactness by variables of cluster objects and also validates the separation, difference by distances between cluster centers.

Table 1
External Clustering Validation Measures
(Aggarwal et al., 2014, Arbelaitz et al., 2013 Halkidi et al., 2002, Halkidi et al., 2002, Liu et al., 2010)

No	External Clustering Validation Measures	Definition
1.	Entropy (*E*)	$-\sum A_i\left(\sum_j \frac{A_{ij}}{A_i}\log\frac{A_{ij}}{A_i}\right)$
2.	Purity (*P*)	$\sum_i A_i\left(max_j \frac{A_{ij}}{A_i}\right)$
3.	F-measure (*F*)	$\sum_j A_i max_i\left[2\frac{\frac{A_{ij}}{A_I}\frac{A_{ij}}{A_I}}{\frac{A_{ij}}{A_i}+\frac{A_{ij}}{A_i}}\right]$
4.	Variation of Information (*VI*)	$\sum_i A_i log_{A_i} - A_j log_{A_j} - 2\sum_i\sum_j A_{ij}\log\frac{A_{ij}}{A_iA_j}$
5.	Mutual Information (*MI*)	$\sum_i\sum_j A_{ij}\log\frac{A_{ij}}{A_iA_j}$
6.	Rand statistic (*R*)	$\frac{\left[\binom{n}{2}-\sum_i\binom{n_{i.}}{2}-\sum_i\binom{n_{.j}}{2}-\sum_{ij}\binom{n_{ij}}{2}\right]}{\binom{n}{2}}$
7.	Jaccard coefficient (*J*)	$\frac{\sum_{ij}\binom{n_{ij}}{2}}{\sum_i\binom{n_{i.}}{2}+\sum_i\binom{n_{.j}}{2}-\sum_{ij}\binom{n_{ij}}{2}}$
8.	Fowlkes & Mallows index (*FM*)	$\frac{\sum_{ij}\binom{n_{ij}}{2}}{\sqrt{\sum_i\binom{n_{i.}}{2}+\sum_i\binom{n_{.j}}{2}}}$
9.	Hubert Γ statistic I (Γ)	$\frac{\left[\binom{n}{2}-\sum_{ij}\binom{n_{ij}}{2}-\sum_i\binom{n_i}{2}-\sum_j\binom{n_{.j}}{2}\right]}{\sqrt{\sum_i\binom{n_{i.}}{2}+\sum_j\binom{n_{.j}}{2}}\left[\binom{n}{2}-\sum_i\binom{n_{i.}}{2}\right]\left[\binom{n}{2}-\sum_j\binom{n_{.j}}{2}\right]}$
10.	Hubert Γ` statistic II (Γ`)	$\frac{\binom{n}{2}-2\sum_i\binom{n_{i.}}{2}-2\sum_j\binom{n_{.j}}{2}-4\sum_{ij}\binom{n_{ij}}{2}}{\binom{n}{2}}$
11.	Minkowski score (*MS*)	$\frac{\sqrt{\sqrt{\sum_i\binom{n_{i.}}{2}+\sum_j\binom{n_{.j}}{2}}-2\sum_{ij}\binom{n_{ij}}{2}}}{\sqrt{\sum_j\binom{n_{.j}}{2}}}$
12.	Classification error (ε)	$1-\frac{1}{n}max_\sigma\sum_j n_\sigma(j)j$
13.	Van Dongen criterion (*VD*)	$\frac{2n-\sum_i max_j n_{ij}-max_j n_{ij}}{2n}$
14.	Micro-average precision (*MAP*)	$\sum_i A_i\left(max_j\frac{A_{ij}}{A_i}\right)$
15.	Goodman-Kruskalcoeff (*GK*)	$\sum_i A_i\left(1-max_j\frac{A_{ij}}{A_i}\right)$

16.	Mirkin metric (*M*)	$\sum_i n_{i.}^2 + \sum_j n_{.j}^2 - 2\sum_i \sum_j n_{ij}^2$

Table 2
Internal Clustering Validation Measures
(Aggarwal et al., 2014, Arbelaitz et al., 2013 Halkidi et al., 2002, Halkidi et al., 2002, Liu et al., 2010)

No	Internal Clustering Validation Measures	Definition
1.	Root-mean-square standard deviation	$\left\{\frac{\sum_i \sum_{x\in Ci} \|\|x - Ci\|\|^2}{[A\sum_i (n_i - 1)]}\right\}^{\frac{1}{2}}$
2.	R-squared (*RS*)	$\frac{\sum_{x\in D} \|\|x - Ci\|\|^2 - \sum_i \sum_{x\in Ci} \|\|x - Ci\|\|^2}{\sum_{x\in D} \|\|x - Ci\|\|^2}$
3.	Modified Hubert Γstatistic (Γ)	$\frac{2}{n(n-1)} \sum_{x\in D} \sum_{x\in D} \sum_{y\in D} d_{(x,y)} d_{x\in Ci, y\in Cj\,(Ci,Cj)}$
4.	Calinski-Harabasz index (*CH*)	$\frac{\sum_i n_i d^2(Ci, C) / NC - 1}{\sum_i \sum_{x\in Ci} d^2(x - Ci) / n - NC}$
5.	*I* index (*I*)	$\left(\frac{1}{NC} . \frac{\sum_{x\in D} d\,(x, c)}{\sum_i \sum_{x\in Ci} d(x,\ Ci)} . max_{ij} d(Ci, Cj)\right)^p$
6.	Dunn's indices (*D*)	$min_i \left\{ min_j \frac{\left(max_{x\in Ci, y\in Cj}\, d(x,y)\right)}{max_k \{max_{x,y\in Ck}\, d(x,y)\}} \right\}$
7.	Silhouette index (*S*)	$\frac{1}{NC} \sum_i \left\{ \frac{1}{n_i} \sum_{x\in Ci} \frac{b(x) - a(x)}{\max[b(x),\ a(x)]} \right\}$
8.	Davies-Bouldin index (*DB*)	$\frac{1}{NC} \sum_i max_{j,j\neq i} \left\{ \left[\frac{1}{n_i} \sum_{x\in Ci} d(x, Ci) + \frac{1}{n_j} \sum_{x\in Cj} d(x, Cj)\right] / d(Ci, Cj) \right\}$
9.	Xie-Beni index (*XB*)	$\sum_i \sum_{x\in Ci} d^2(x, Ci) / n. min_{i,j\neq i} d^2(Ci, Cj)$
10.	SD validity index (*SD*)	$Dis(NC_{max}) Scat(NC) + Dis(NC)$ $Scat(NC) = \frac{1}{NC} \sum_i \|\|\sigma(Ci)\|\| / \|\|\sigma(D)\|\|$ $Dis(NC) = \frac{max_{ij}\, d(Ci, Cj)}{min_{ij}\, d(Ci, Cj)} \sum_i \left(\sum_j d(Ci, Cj)\right)^{-1}$

V CONCLUSION

This paper reviews the core idea of big data, big data mining, big data storage structures, clustering and its taxonomy with the validation measures taxonomy. This paper also defined how to validate the traditional clustering taxonomy by using internal and external validation measures under big data mining. These validation measures define clustering algorithm given a more accurate and reliable cluster under high volume, heterogeneous data and sources with scalability. The first and second section of this paper gives the basic background of evaluation of traditional data to big data, big data dimensions, big data mining, and clustering approach and define how to clustering approach scalable under the big data mining. The third section presents the concept of all existing clustertaxonomies by their creation process

and behaviors. The fourth section defines various cluster validation measures for cluster accuracy under the cluster internal and external natures. The overall this paper presents various clustering approaches for the availabilityof big data mining and their validation approach for cluster reliabilities and accuracy.

REFERENCES

[1] Aggarwal, C. C., & Reddy, C. (2014). Data Clustering Algorithms and Applications. CRC Press Taylor & Francis Group.ISBN 978-1-4665-5822-9.

[2] Arbelaitz, O., Gurrutxaga, I., Muguerza, J., Pérez, J. M., &Perona, I. (2013). An extensive comparative study of cluster validity indices. Pattern Recognition, 46(1), 243-256. doi:10.1016/j.patcog.2012.07.021.

[3] Comprehensive Survey on Distance/Similarity Measures between Probability Density Functions. INTERNATIONAL JOURNAL OF MATHEMATICAL MODELS AND METHODS IN APPLIED SCIENCES, 4(1), 300-307. doi:10.1109/icpr.2000.906010.

[4] Chen, W., Oliverio, J., Kim, J. H., & Shen, J. (2018). The Modeling and Simulation of Data Clustering Algorithms in Data Mining with Big Data. Journal of Industrial Integration and Management, 12(4), 1-16. doi:10.1142/s2424862218500173.

[5] Chong, D., & Shi, H. (2015). Big data analytics: A literature review. Journal of Management Analytics, 2(3), 175-201. doi:10.1080/23270012.2015.1082449.

[6] Fahad, A., Alshatri, N., Tari, Z., Alamri, A., Khalil, I., Zomaya, A. Y., . . .Bouras, A. (2014). A Survey of Clustering Algorithms for Big Data: Taxonomy and Empirical Analysis. IEEE Transactions on Emerging Topics in Computing, 2(3), 267-279. doi:10.1109/tetc.2014.2330519.

[7] Gan, G., Ma, C., & Wu, J. (2007). Data clustering: Theory, algorithms, and applications. Philadelphia, PA: SIAM, Society for Industrial and Applied Mathematics.

[8] Gandomi, A., &Haider, M. (2015). Beyond the hype: Big data concepts, methods, and analytics. International Journal of Information Management, 35(2), 137-144. doi:10.1016/j.ijinfomgt.2014.10.007.

[9] Halkidi, M., Batistakis, Y., &Vazirgiannis, M. (2002). Cluster validity methods. ACM SIGMOD Record, 31(2), 40. doi:10.1145/565117.565124.

[10] Kocher, M., & Savoy, J. (2017). Distance measures in author profiling. Information Processing & Management, 53(5), 1103-1119. doi:10.1016/j.ipm.2017.04.004.

[11] Lee, I. (2017). Big data: Dimensions, evolution, impacts, and challenges. Business Horizons, 60(3), 293-303. doi:10.1016/j.bushor.2017.01.004.

[12] Liu, Y., Li, Z., Xiong, H., Gao, X., & Wu, J. (2010). Understanding of Internal Clustering Validation Measures. 2010 IEEE International Conference on Data Mining. doi:10.1109/icdm.2010.35.

[13] Manning, C. D. ,Raghavan, P. , &Schütze, H. (2008). Introduction to information retrieval . Cambridge: Cambridge University Press .

[14] Oussous, A., Benjelloun, F., Lahcen, A. A., &Belfkih, S. (2018). Big Data technologies: A survey. Journal of King Saud University - Computer and Information Sciences, 30(4), 431-448. doi:10.1016/j.jksuci.2017.06.001.

[15] Pandove, D., &Goel, S. (2015). A comprehensive study on clustering approaches for big data mining. In Proceedings of IEEE 2nd International Conference on Electronics and Communication Systems (pp. 1333-1338). IEEE Xplore Digital Library. doi:10.1109/ecs.2015.7124801.

[16] Sardar, T. H., & Ansari, Z. (2018). An analysis of MapReduce efficiency in document clustering using parallel K-means algorithm. Future Computing and Informatics Journal, 3(2), 200-209. doi:10.1016/j.fcij.2018.03.003.

[17] Shirkhorshidi, A. S., Aghabozorgi, S., Wah, T. Y., &Herawan, T. (2014). Big Data Clustering: A Review. In Murgante B. et al. (eds), International Conference on Computational Science and Its Applications (Vol. 8583, Lecture Notes in Computer Science, pp. 707-720). Springer. doi:10.1007/978-3-319-09156-3_49.

[18] Siddiqa, A., Hashem, I. A., Yaqoob, I., Marjani, M., Shamshirband, S., Gani, A., &Nasaruddin, F. (2016). A survey of big data management: Taxonomy and state-of-the-art. Journal of Network and Computer Applications, 71, 151-166. doi:10.1016/j.jnca.2016.04.008.