International Journal of Contemporary Computer Research (IJCCR), Vol.2 Issue.1 (March, 2018)
ISSN: 2600-9048

Review: Data Clustering Approaches

Mohd Shaifulnizam Shafii[1], Wael M. S. yafooz[2]

[1] *Faculty of Computer and Information Technology, Al-Madinah International University, Malaysia, fibonac32@gmail.com*

Received 09 January 2018; accepted 19 March 2018

Abstract

Data clustering is part of data mining. Data mining is the process extracting useful information, patterns and trends from collective of raw data. The raw data have hidden pattern and information that are useful to retrieve for needed application and studies. The clustering is the efficient technique of data mining which will cluster the similar and dissimilar type of data. The clustering techniques are of many types like density based, hierarchal clustering, partitional clustering, grid based and more on. In this paper, various techniques of clustering and their parameter have been reviewing.

Keywords: Data mining, clustering, portioning.

1. Introduction

Clustering is group of data with similar object and different objects are grouped differently according to the object similarity. Division of data is according to group of similar objects. It is also called unsupervised classification. There are various techniques partition Based- K means, K-Medoils, K-modes, PAM, CLARANS, CLARA, and FCM, hierarchical Based – CURE, BIRCH, ROCK, Echidna, and Chameleon, grid Based- STING, CLIQUE, Wave Cluster, Optigrid, density Based-DBSCAN, OPTICS, DBCLASD, DENCLUE and model Based-COBWEB, EM, CLASSIT,SOMs. Clustering algorithm selection is depending on the data set and final application of the clustered data discovery. Clustering is in several areas such as statistical data analysis, machine learning, pattern recognition and image analysis [1].

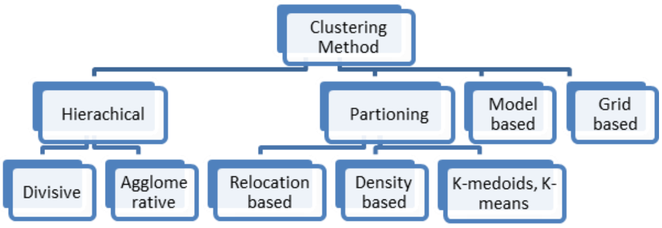

Figure 1: Clustering methods

Hierarchical algorithms build clusters gradually and partitioning algorithms learn clusters directly. Both algorithms attempts to discover clusters by iteratively relocating points between subsets or try to identify clusters as areas highly populated with data [2].

Hierarchical
Build clusters gradually while grow up
Partitioning
Learn clusters directly

Figure 2: Hierarchical and Partitioning Clustering

Data mining with clustering end to end process as figure 3 explained below:

- Data Collection: Extraction of relevant data objects from the requirement data sources.
- Data Cleaning: It is done after extraction of data from the source.
- Representation: Proper preparation of the data to become suitable for the clustering algorithm.
- Clustering Tendency: Checks whether the data in hand has a natural tendency to cluster or not.
- Clustering Strategy: Makes choice of clustering algorithm and initial parameters.
- Validation: Based on manual examination and visual techniques.
- Interpretation: Clustering results interpretations are combined with other studies for further analysis and analytic studies [3].

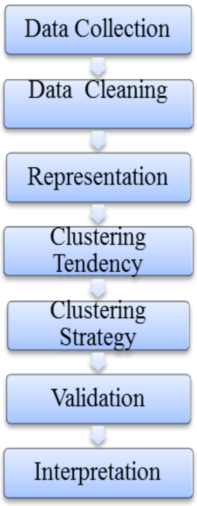

Figure 3: Data mining using clustering process end to end

2. Hierarchical Clustering

Hierarchical clustering is to select a distance measure. A simple measure is Manhattan distance, equal to the sum of absolute distances for each variable. A more common measure is Euclidean distance, computed by finding the square of the distance between each variable, summing the squares, and finding the square root of that sum [4.

The Manhattan distance function computes the distance that would be travelled to get from one data point to the other if a grid-like path is followed. The Manhattan formula distance between a point X=(X1, X2...) and a point Y= (Y1, Y2...):

$$d = \sum_{i=1}^{n} \left| X_i - Y_i \right|$$

The Euclidean formula distance is between a point X (X1, X2...) and a point Y (Y1, Y2...):

$$d = \sqrt{\sum_{j=1}^{n} (x_j - y_j)^2}$$

There are two hierarchical which are agglomerative and divisive [5].

A. Agglomerative

Agglomeration follows at a better distance between clusters than the previous agglomeration, we can decide to stop clustering whichever when the clusters are too far apart to be merged or when a sufficiently small number of cluster are available. Agglomeration is bottom-up as figure 3. Basic agglomerative hierarchical clustering algorithm:

1: Compute the proximity matrix, if necessary.
2: Repeat
3: Merge the closest two clusters.
4: Update the proximity matrix to reflect the proximity between the new cluster and the original clusters [6].
5: Until only one cluster remains.

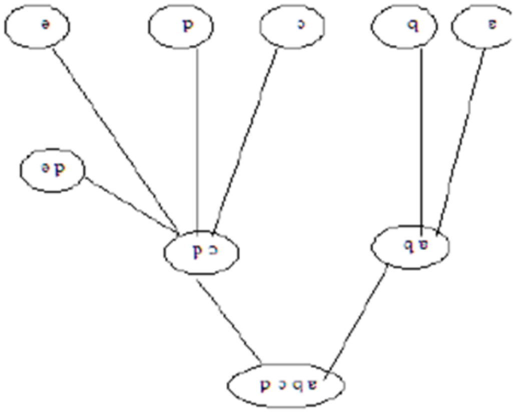

Figure 4: Agglomerative hierarchical clustering

B. Divisive clustering

Top-down clustering is called divisive clustering. The cluster is split using a flat clustering algorithm. Top- down clustering is need a second, flat clustering algorithm as a “subroutine”. The advantage of being more efficient if we do not generate a complete hierarchy all the way down to individual document leaves.

For a fixed number of top levels, using an efficient flat algorithm like K-means, top-down algorithms are linear in the number of documents and clusters [7].

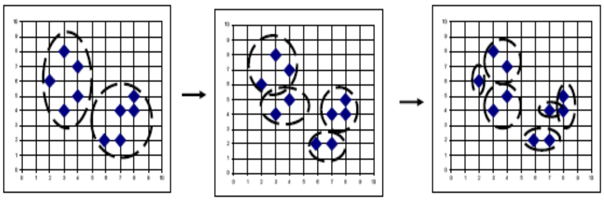

Figure 5: Divisive top-down

3. Partitional Clustering

A. K-means algorithm

K-means algorithm assigns each point to the cluster whose center also called centroid is nearest. The center is the average of all the points in the cluster that is, its coordinates are the arithmetic mean for each dimension separately over all the points in the cluster. The pseudo code of the k-means algorithm is to explain how it works:

A. Choose K as the number of clusters.

B. Initialize the codebook vectors of the K clusters (randomly, for instance)

C. For every new sample vector:

a. Compute the distance between the new vector and every cluster's codebook vector.

b. Re-compute the closest codebook vector with the new vector, using a learning rate that decreases in time [8].

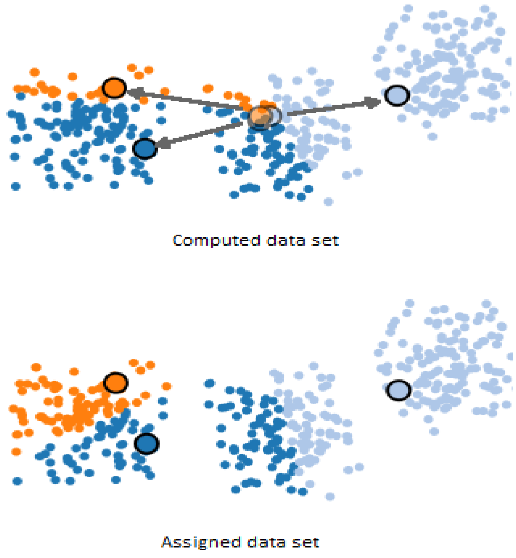

Figure 6: K-means algorithm

B. K-medoids algorithm

K-medoids algorithm is each cluster is represented by one of the objects located near the center of the cluster. The pseudo code of the k-medoids algorithm is as follow, arbitrarily choose k objects as the initial medoids Repeat Assign each remaining object to the cluster with the nearest medoids Randomly select a non-medoid object Orandom. Compute the total cost, S, of swapping Oj with Orandom. If S<0 the swap Oj with Orandom to form the new set of k-medoids, until no changes [9].

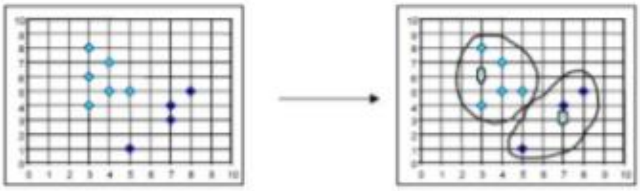

Figure 7: K-medoids algorithm

C. Density based

Density based clustering algorithms are devised to discover arbitrary-shaped clusters. In this approach, a cluster is regarded as a region in which the density of data objects exceeds a threshold. Density based algorithm is DBSCAN and SSN [10].

The DBSCAN algorithm depends on a density-based notion of clusters. Clusters are identified by looking at the density of points. Regions with a high density of points depict the existence of clusters whereas regions with a low density of points indicate clusters of noise or clusters of outliers. This algorithm is particularly suited to deal with large datasets, with noise, and is able to identify clusters with different sizes and shapes. The key idea of the DBSCAN algorithm is that, for each point of a cluster, the neighborhood of a given radius has to contain at least a minimum number of points, that is, the density in the neighborhood has to exceed some predefined threshold. This algorithm needs three input parameters: - k, the neighbor list size; - Eps, the radius that delimitate the neighborhood area of a point (Eps neighborhood); - MinPts, the minimum number of points that must exist in the Eps-neighborhood [11].

The clustering process is based on the classification of the points in the dataset as core points, border points and noise points, and on the use of density relations between points to form the clusters. The pseudo code of the DBSCAN algorithm is to explain how it works: To clusters a dataset, our DBSCAN implementation starts by identifying the k nearest neighbours of each point and identify the farthest k nearest neighbour. The average of all this distance is then calculated. For each point of the dataset the algorithm identifies the directly density-reachable points using the Eps threshold provided by the user and classifies the points into core or border points. It then loop trough all points of the dataset and for the core points it starts to construct a new cluster with the support of the GetDRPoints() procedure that follows the definition of density reachable points. In this phase the value used as Eps threshold is the average distance calculated previously. At the end, the composition of the clusters is verified in order to check if there exist clusters that can be merged together [12].

SNN algorithm measure the density is defined as the sum of the similarities of the nearest neighbours of a point. Points with high density become core points, while points with low density represent noise points. All remainder points that are strongly similar to a specific core points will represent a new clusters. The SNN algorithm needs three inputs parameters: - K, the neighbours‟ list size; - Eps, the threshold density; - MinPts, the threshold that define the core points. The pseudo code of the SSN algorithm is to explain how it works: Define the input parameters. Find the K nearest neighbours of each point of the dataset. Then the similarity between pairs of points is calculated in terms of how many nearest neighbors the two points share. Using this similarity measure, the density of each point can be calculated as being the numbers of neighbors with which the number of shared neighbors is equal or greater than Eps. The points are classified as being core points, if the density of the point is equal or greater than MinPts. At this point, the algorithm has all the information needed to start to build the clusters. Those start to be constructed around the core points. However, these clusters do not contain all points. They contain only points that come from regions of relatively uniform density. The points that are not classified into any cluster are classified as noise points [13].

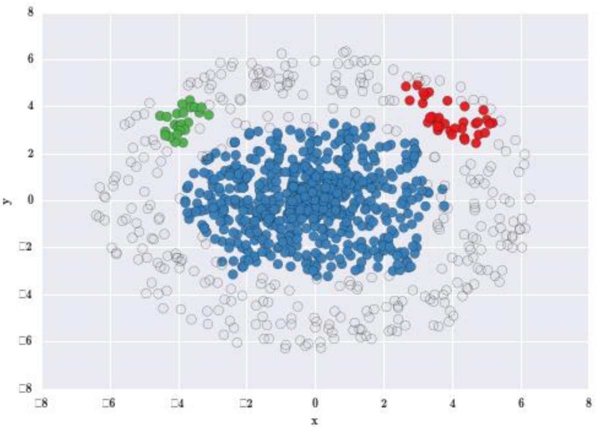

Figure 8: Density based model

4. Grid Based Algorithm

The grid based clustering approach uses a multiresolution grid data structure. It quantizes the space into a finite number of cells that form a grid structure on which all the operations for clustering are performed. Grid approach includes STING (Statistical Information Grid) approach and CLIQUE Basic.

1. Define a set of grid-cells
2. Assign objects to the appropriate grid cell and compute the density of each cell.
3. Eliminate cells, whose density is below a certain threshold t.
4. Form clusters from contiguous (adjacent) groups of dense cells [14].

The pseudo code of the STING algorithm is to explain how it works: The spatial area is divided into rectangular cells. There are several levels of cells corresponding to different levels of resolution. Each cell is partitioned into a number of smaller cells in the next level. Statistical info of each cell is calculated and stored beforehand and is used to answer queries Parameters of higher level cells can be easily calculated from parameters of lower level cell count, mean, s, min, max type of distribution—normal, uniform, etc [15]. Use a top-down approach to answer spatial data queries Start from a pre-selected layer—typically with a small number of cells from the pre-selected layer until you reach the bottom layer do the following: For each cell in the current level compute the confidence interval indicating a cells relevance to a given query: [16]

1. If is relevant, include the cell in a cluster
2. If it irrelevant, remove cell from further consideration
3. Otherwise, look for relevant cells at the next lower layer

CLIQUE (Clustering in QUEst) is a bottom-up subspace clustering algorithm that constructs static grids. . CLIQUE is a density and grid based. Clustering process in CLIQUE as follow: [17]

1. CLIQUE partitions the d- dimensional data space into non-overlapping rectangular units called grids according to the given grid size and then find out the dense region according to a given threshold value. A unit is dense if the data points in this are exceeding the threshold value.
2. Clusters are generated from the all dense subspaces by using the apriori approach. CLIQUE algorithm generates minimal description for the clusters obtained by first determining the maximal dense regions in the subspaces and then minimal cover for each cluster from that maximal region. It repeats the same procedure until covered all the dimensions [18].

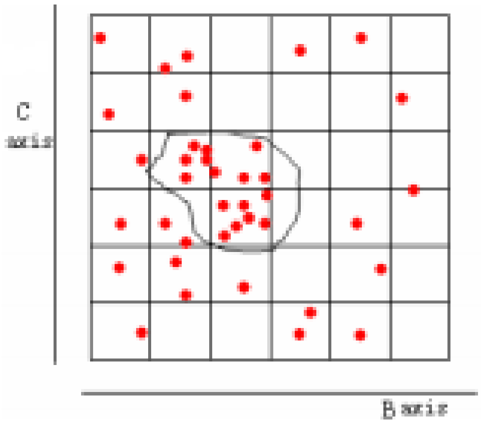

Figure 9: Grid based cluster as in circle

5. MODEL BASED

Model based method is trying to optimize the fit between the given data and some mathematical model. Such methods often based on the assumption that the data are generated by mixture of underlying probability distributions. Model-Based Clustering methods follow two major approaches: Statistical Approach or Neural network approach [19].

1. Clustering is also performed by having several units competing for the current object.
2. The unit whose weight vector is closest to the current object wins.
3. The winner and its neighbors learn by having their weights adjusted.
4. SOMs are believed to resemble processing that can occur in the brain.
5. Useful for visualizing high-dimensional data in 2D- or 3D space. [20]

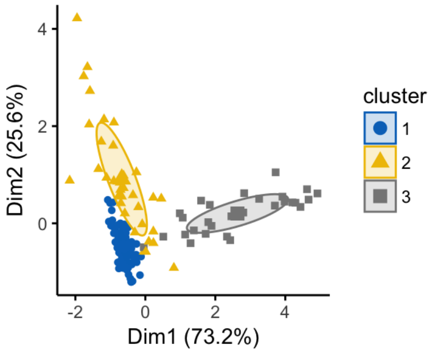

Figure 10: Model based cluster plot

6. Discussion

The paper review different clustering techniques. There are several consideration need to look for choosing suitable clustering techniques such as assessment of results, choice of appropriate number of clusters, data preparation, proximity measures and handling outliers.

Assessment of results

The data mining clustering process starts with the assessment of whether any cluster tendency has a place at all, and correspondingly includes, appropriate attribute selection, and in many cases feature construction. It finishes with the validation and evaluation of the resulting clustering system. The clustering system can be assessed by an expert, or by a particular automated procedure. Traditionally, the first type of assessment relates to two issues: (1) cluster interpretability, (2) cluster visualization. Interpretability depends on the technique used. Model based probabilistic algorithms have better scores in this regard. K-means and k-medoid methods generate clusters that are interpreted as dense areas around centroids or medoids and, therefore, also score well.

Regarding automatic procedures, when two partitions are constructed (with the same or different number of subsets k), the first order of business is to compare them. Sometimes the actual class label s of one partition is known. Still clustering is performed generating another label j. The situation is similar to testing a classifier in predictive mining when the actual target is known. Comparison of s and j labels is similar to computing an error, confusion matrix, and more on in predictive mining.

Choice of appropriate number of clusters:

In many methods number k of clusters to construct is an input user parameter. Running an algorithm several times leads to a sequence of clustering systems. Each system consists of more granular and less-separated clusters. In the case of k-means, the objective function is monotone decreasing. Therefore, the answer to the question of which system is preferable is not trivial. Many criteria have been introduced to find an optimal k. For instance, a distance between any two centroids (medoids) normalized by corresponding cluster radii (standard deviations) and averaged (with cluster weights) is a reasonable choice of coefficient of separation.

Data preparation:

Attributes transformation and clustering have already been discussed in the context of dimensionality reduction. The practice of assigning different weights to attributes and/or scaling of their values is widespread and allows constructing clusters of better shapes. In real-life applications it is crucial to handle attributes of different nature. For example, astronomical images are characterized by color, wavelength, texture, shape, and location, resulting in five attribute subsets. Some algorithms depend on the effectiveness of data access. To facilitate this process data indices are constructed. Index structures used for spatial data, include KD-trees. A blend of attribute transformations (DFT, Polynomials) and indexing technique is present in many methods. The major application of such data structures is in nearest neighbors search.

Proximity Measures:

Both hierarchical and partitioning methods use different distances and similarity measures. The usual distance Lp:

$$Lp = d(x,y) = ||x-y||_{1\leq p\leq \propto}$$

In which lower p corresponds to a more robust estimation (therefore, less affected by outliers). Euclidean (p=2) distance is by far the most popular choice used in k-means objective function (sum of squares of distances between points and centroids) that has a clear statistical meaning of total inter-clusters variance. The distance that returns the maximum of absolute difference in coordinates is also used and corresponds to $p = \infty$.

Handling Outliers:Applications that derive their data from measurements have an associated amount of noise, which can be viewed as outliers. Alternately, outliers can be viewed as records having abnormal behaviour. In general, clustering techniques do not distinguish between the two: neither noise nor abnormalities fit into clusters. Correspondingly, the preferable way to deal with outliers in partitioning the data is to keep one extra set of outliers.

There are multiple ways of how descriptive learning handles outliers. If a data pre-processing phase is present, it usually takes care of outliers. For example, this is the case with grid-based methods. They simply rely on input thresholds to eliminate low-populated cells. Other algorithms revisit outliers during the decision tree rebuilds, but in general handle them separately, by producing a partition plus a set of outliers. Certain algorithms have specific features for outliers handling. Some of them use shrinkage of cluster representatives to suppress the effects of outliers. K-medoids methods are generally more robust than k-means methods with respect to outliers. Others (such as DBSCAN) use concepts of internal (core), boundary (reachable), and outliers (non-reachable) points. The algorithm CLIQUE goes a step further: it eliminates subspaces with low coverage.

Table1: Summarized Comparison of Clustering Techniques

Clustering Techniques	Method	Algorithm
Partitioning	Construct k partitions (k <= n) and then evaluate them by some criterion, minimizing the sum of square errors Each group has at least one object, each object belongs to one group Iterative Relocation Technique Avoid Enumeration by storing the centroids	k-means, k-medoids, CLARANS
Hierarchical	Create a hierarchical decomposition of the set of data using some criterion Agglomerative Vs Divisive Rigid – Cannot undo Perform Analysis of linkages Integrate with iterative relocation	Diana, Agnes, BIRCH
Density Based	Distance based methods – Spherical Clusters Density – For each data point within a given cluster the neighborhood should contain a minimum number of points	DBSCAN, OPTICS
Grid Based	Object space – finite number of cells	STING,

	forming grid structure and fast processing time	WaveCluster, CLIQUE
Model-based	A model is hypothesized for each of the clusters and tries to find the best fit of that model to each other	EM, COBWEB

7. Conclusion

The review in this paper shows data clustering techniques are various. Clustering algorithms can be categorized into partition, hierarchical, density based and grid based algorithms. They are from simple and complex algorithm. There is not perfection in one algorithm, they all have own advantages and disadvantages. Clustering is all about group, and there is need a group in the algorithm. By applying a cluster analysis we are hypothesizing that the groups is exist.

Acknowledgment

Thank you to faculty of Computer and Information Technology, Al-Madinah International University to make the data mining journal learning and publication process is possible.

References

[1] Ajit Kumar, Dharmender Kumar and S. K. Jarial, “Review on Artificial Bee Colony Algorithms and Their Applications to Data Clustering,” Bulgarian academy of sciences, Cybernetics and information technologies Volume 17, No 3, 2017.

[2] Anjana Gosaina and Sonika Dahiyab, “Performance Analysis of Various Fuzzy Clustering Algorithms: A Review,” 7th International Conference on Communication, Computing and Virtualization 2016,2016.

[3] El-Ebiary, Y. (2015). 37-Data Analysis Techniques. Al-Madinah Technical Studies| -مجلة جامعة المدينة العالمية للعلوم التقنية 1 ,ماليزيا(3).

[4] Anshul Yadav and Sakshi Dhingra, “A review on k-means clustering technique,” International Journal of Latest Research in Science and Technology ISSN (Online):2278-5299 Volume 5, Issue 4: Page No.13-16, July - August 2016.

[5] Mr. Mukesh K. Deshmukh and Prof. A. S. Kapse, “A Survey On Outlier Detection Technique In Streaming Data Using Data Clustering Approach,“International Journal Of Engineering And Computer Science ISSN: 2319-7242 volume 5 Issue 1 January 2016.

[6] Qasem a. al-radaideh, Adel abu assaf and Eman alnagi, “Predicting stock prices using data mining techniques,”, The International Arab Conference on Information Technology (ACIT’2013), 2013.

[7] Anju and Preeti Gulia,“Enhancement in K-Mean Clustering in Big Data,” International Journal of Advanced Research in Computer Science Volume 8, No. 3, April 2017.

[8] Amandeep Kaur Mann and Navneet Kaur, “Review Paper on Clustering Techniques,” Global Journal of Computer Science and Technologym Software & Data Engineering Volume 13 Issue 5 Version 1.0, 2013.

[9] Rashmi P. Dagde and Snehlata Dongre, “A Review on Clustering Analysis based on Optimization Algorithm for Datamining,” IJCSN International Journal of Computer Science and Network, Volume 6, Issue 1, February 2017.

[10] Meenu Sharma and Mr. Kamal Borana, “Clustering In Data Mining : A Brief Review,” International Journal Of Core Engineering & Management (IJCEM) Volume 1, Issue 5, August 2014.

[11] Asha Devi and Saurabh Sharma, “Review on Analysis of Clustering Techniques in Data Mining,”. ISSN (e): 2250 – 3005 || Volume, 07 Issue, 08, August 2017.

[12] Brinda Gondaliya, “Review paper on clustering techniques,” International Journal of Engineering Technology, Management and Applied Sciences, Volume 2 Issue 7, December 2014.

[13] A. Fahad, N. Alshatri, Z. Tari, Member, IEEE , A. Alamri, I. Khalil A. Zomaya, Fellow, IEEE, S. Foufou, and A. Bouras, “A Survey of Clustering Algorithms for Big Data: Taxonomy & Empirical Analysis,” IEEE Transactions on Emerging Topics in Computing, 2014.

[14] Jyoti, Neha Kaushik and Rekha, “Review paper on clustering and validation techniques,” IJRASET, Vol. 2, Issue V, May 2014.

[15] Mamta Mor,“A Review on Various Clustering Techniques in Data Mining,” Mamta Mor, International Journal of Computer Science & Communication Networks,Vol 6(3),138-142,2016.

[16] [Jyoti Yadav and Monika Sharma, “A Review of K-mean Algorithm,” International Journal of Engineering Trends and Technology (IJETT) – Volume 4 Issue 7- July 2013.

[17] Shraddha Shukla and Naganna S., “A Review ON K-means DATA Clustering APPROACH,” International Journal of Information & Computation Technology, ISSN 0974-2239 Volume 4, Number 17, 2014.

[18] Amandeep Kaur Mann and Navneet Kaur, “Survey Paper on Clustering Techniques,” International Journal of Science, Engineering and Technology Research (IJSETR) Volume 2, Issue 4, April 2013.

[19] Jamal Uddin, Rozaida Ghazali and Mustafa Mat Deris, “An Empirical Analysis of Rough Set Categorical Clustering Techniques,” PLOS ONE | DOI:10.1371/journal.pone.0164803 January 9, 2017.

[20] Antonio M. Ortiz, Member, IEEE, Dina Hussein, Soochang Park, Member, IEEE, Son N. Han, Student Member, IEEE, and Noel Crespi, Senior Member, IEEE, , “The Cluster Between Internet of Things and Social Networks: Review and Research Challenges,” IEEE internet of things journal, vol. 1, no. 3, june 2014.

[21] Yafooz, W. M., Abidin, S. Z., Omar, N., & Hilles, S. (2016, September). Interactive Big Data Visualization Model Based on Hot Issues (Online News Articles). In International Conference on Soft Computing in Data Science (pp. 89-99). Springer, Singapore.