THINK INDIA JOURNAL ISSN: 0971-1260

Vol-22-Issue-3-July-2019

Framework To Improve The Processing of The Asymmetric Clustering

Sheveta Vashisht[1],Mehak[2]

[1]Assistant Professor, School of Computer Science & Engineering, Lovely Professional University, Punjab, India.
[2]Assistant Professor, School of Computer Science & Engineering, Lovely Professional University, Punjab, India.

Abstract

With the rapid development of information technology an endless amount of data and information is available and more emphasis is paid on its analysis. Data mining helps in extracting useful insights from large and detailed collections of data. Data mining currently has attention within the business related to information security due to the availability of the large amount of data and rapid growth in the size of data there exist the need to do the analysis of this data to extract useful patterns which can befurther used in number of fields. Different clustering calculations have been created to bunch information into clusters in various domains. In any case, these clustering calculations work successfully either on immaculate numeric information or on immaculate categorical information; most of them perform ineffectively on blended categorical and numeric information sorts. There are number of information mining methods and asymmetric clustering is one of them. Asymmetric strategy is sort of unsupervised learning. In order to improve neural network accuracy and to reduce processing time andup gradation is done in asymmetric clustering algorithm.

Keywords: Clustering, Asymmetric, Noise, Escape time, Accuracy.

I. INTRODUCTION

Data Mining is utilized on huge information set to extricate designs and information from it. The data which is gotten from information mining is visualized within the frame of graphcharts, pie charts and tables. Knowledge discovery database is another name for data mining. In order to find relationship between various numberof fields frequent item set is used in data mining [5].In order to find the frequent data set item association rules are used.

 Copyright © 2019Author

This concept of association rules is used is utilized in the different areas like retail stores, showcase technique and stock showcase and other areas also.

There are numerous strategies in which are utilized in data mining. These strategies are:

1. Classification: The most basic and oldest technique to process data is which is supported by expert system. Primarily classification is employed during a set of information to classify every item into one of the predefined set of categories or teams. Mathematical techniques such as decision trees, call trees, neural network, applied math and statistics are used for classification [3]

2. Association: In data mining association is one of the techniques which ismost openly used. Association data mining technique is also called as relation technique because on the basis of relationship between items patterns are discovered. In market basket analysis customer's behaviour is judged by observing the items which customer's buys frequently. Association techniques are used to research customer's habit in buying items. On the basis of past records salesman can attract customers by giving them combo offers on items which they mostly buys.

3. Prediction: The prediction is the technique of a knowledge mining that confirms association among independent variables and relationship between dependent and independent entities. It's used to predict the profit of the corporate by analysing its history record. It's based mostly upon dependent and independent variables each. Then supported the historical sale and profit information, a fitted regression line that's used for profit prediction is drawn.[9].

4. Clustering: Clustering is a technique in objects of similar category is placed in one group and other are in different group using automatic technique [10].

 Copyright © 2019Author

ISSN: 0971-1260

Vol-22-Issue-3-July-2019

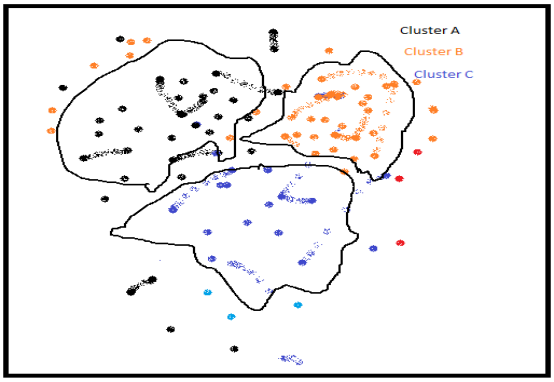


Figure.1 Clustering in Data Mining

1.1. Clustering in Data Mining

Clustering technique defines categories and place objects that are associated with them in one category whereas classification objects are placed in already defined categories. So clustering places the objects that have same properties into single cluster and objects that have different properties into other [7].

If the value of the objects is more than certain defined threshold value that are kept in one category of the cluster and other objects having value less than threshold that object is placed in other cluster which is according to the similarity property of the object.

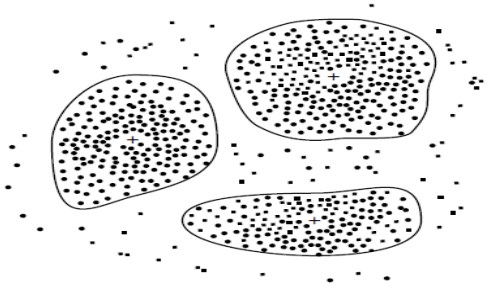


Figure 1.2 Cluster and Outlier

In above figure illustrate the objects ,the objects which are not inside the cluster represent the outliners and objects with the same property are inside the cluster in groups.The main difference between clustering and the nearest neighbor technique is that clustering is an unsupervised learning technique on the other hand nearest neighbor is generally used for

Copyright © 2019Author

ISSN: 0971-1260

Vol-22-Issue-3-July-2019

prediction or a supervised learning technique. Unsupervised learning methods are not beneath a few supervision though administered strategy is beneath supervision like learning beneath instructor. In expectation, the designs that are found within the database and displayed within the model are continuously the foremost important patterns within the database for performing a few specific forecast. In clustering there's no specific sense of why certain records are close to each other or why they all drop into the same cluster.

1.2 Data Mining techniques for Clustering: Clustering is an unsupervised learning technique. There are many techniques for clustering in data mining. These are as follow:

1.2.1 Partitioning Clustering: In partitioning clustering the partition of objects is done and is called clusters. The clusters formed to optimize an objective partitioning criterion in way that on the basis of distance the dissimilarity function, the similar objects are in cluster and the objects of different cluster are not similar with respect to attributes of data set. There are two partitioning methods one is k-means and other is k-medoids.

1.2.2 Clustering based on density:

The base of the maximum partitioning based algorithms is the distance that exists between the objects. But by using these methods it's difficult to find arbitrary shaped clusters and spherical clusters are identified easily. So to overcome the identification issue of the arbitrary clusters ,a density based new methods is used. This new methods has the basis of clustering based on the concept of density .In this procedure the groups grows in size depending on the density of neighbouring objects as and when they exceed some pre specified threshold values[3].

1.2.3 Grid Based Clustering: In clustering which is grid dependent, the data structure used is multitier. It utilizes space of object in the finite number of cells which makes this structure on which all clustering operations are performed. This methodology has quick preparing time and autonomous of number of information items however it is subject to the quantity of data cells in the defined dimension of space. STINGS, CLIQUE are the examples of grid based methods. [8].

Copyright © 2019Author

II. LITERATURE REVIEW

In paper [1] they have got noted that mining arbitrary of clusters which are in arbitrary shape from massive data sets is a challenge in data mining. Many high time based complexity algorithms have been proposed as the solution to this problem. Calculation cost of such algorithms can be saved by using the algorithms which reduces the data set to smaller amount of the representative data. But the clustering performance can be affected by the user defined reduced ratios.CLASP a powerful set of rules for mining arbitrary form clusters algorithm which is presented in this paper is an algorithm which successfully preserve the form of statistics of cluster by automatically reducing the size of the records at the time of representation of data set example. At that point to improve the intrinsic relation among the representative data set and to make structures of cluster distinct and more clear for dividing them into clusters the alterations are made in the positions of the representative data. In the end, for identification of the cluster is done by using agglomerative technique of clustering along with PK metric technique of K-nearest based metric. Extensive experiments on each synthetic and real record devices are accomplished, and the effects demonstrate the effectiveness and performance of this technique

In paper [2] they have explained algometric algorithms of hierarchal clustering using the asymmetric similarity measures. There are linkage methods proposed into this research are of two types .In the first type of linkage method that follows bottom up approach the similarity between objects is first measured and then the similarity in the clusters is measured and on the other hand the second type uses the principal of top-down approach in which similarity is measured among the clusters before measuring the similarity in the objects. The tree diagram structure used to show result of hierarchical clustering called dendrogram result of the hierarchical clustering sometimes shows in reversely. At the end, algometric hierarchal clustering and another method top-down approach use hypothesis.

In paper [3] for detection of the noisy instances , Qclean Noise which is a new algorithm is introduced by the author.In this discussion regarding three different supervised technique based algorithm is done.The effects of LDA algorithm ,along with KNN and RPPART which are decision tree classifier are also discussed.In the experntation carried out in this paper it is demonstrated that above mentioned algorithms and prociedure are better in performance then

 Copyright © 2019Author

the ones which are already exisiting and this has been concluded by comparing the Qcleanalgorithm and other mentioned procedures and algorithms on the well-known machine learning's four data sets.

In paper [4] a new method based on principal of hierarchical approach is discussed. According to this approach in order to preserve the asymmetry of the data as in the output obtained the hierarchical clustering is done over the asymmetric network and the output of such network with respect to desirable properties is equal to the ultra-metric space of the finite quasi .The author proves in this paperthat quasi-clustering is the only modified algorithm for single linkage . In addition, the proposed method shows stability and fulfills the invariance properties it has established. New algorithms are being developed and a study of internal migration within the United States illustrates the value of quasi-clustering analysis.

In paper [5] the author gave an outline of data mining and areas where it could be used. As per authors mining could be used to extract information from very large quantities of data. They listed data mining techniques: decision tree and , various techniques of classification methods and nonlinear regression, etc. Author spoke about areas where data mining can be performed to obtain information that can be used to make decisions.

In paper [6] it is discussed that mining is one of the fast growing fields in the computer science.in data mining new patterns are discovered in large datasets. There are many data mining algorithms are developed in last number of years. Single algorithm may not apply on all applications due to the reasons of the different data types of applications. That's why right selection of algorithms does not depends upon the application used but it depends upon the type of data type is used in the applications.

III ASYMMETRIC CLUSTERING

In an asymmetric clustering, the walls may be affirmed strongly and commonly run on a single cluster at a time. The venture which is related to a suitable cluster and it can be routed to that cluster. The Asymmetric Clustering applications are used in electronic trading structures that are utilized by banks [12]. Asymmetric clustering architecture is contrary to the standard stateless server farm where the whole software is simulated across machines. In an asymmetric cluster, business common sense is dividing into walls, wherein each partition can be the singular accent of a set of underlying records [13]. As a resultant, in each node

 Copyright © 2019Author

ISSN: 0971-1260

Vol-22-Issue-3-July-2019

inside the cluster implementing its personal local cache ensuing in high performance analysing and writing without the need to keep a distributed cache between cluster nodes. An uneven cluster is practically the other of a symmetric cluster:

1. At any point when the application is running the partitions can be named and announced.

2. The partitions are exclusive as they run on a single cluster member and are available as singletons

3. The cluster member that is hosting the partition routed to incoming work for a partition.

IV PROPOSED METHDOLOGY

1. Declare rows, columns, integration and load dataset: - This is the first set of algorithm in which the number of rows and columns are defined for the dataset. The second condition is defined to define number of iteration to define cluster quality. In the step of the flowchart the dataset will be loaded to perform clustering operation.

2. Register Clustering Method: - To register clustering method is the second step of the flowchart in which we defined the two clustering method. The first method is K-mean clustering and second method is asymmetric clustering. According to the selected method the operation of clustering will be performed on the dataset.

3. Apply K-mean Normalization and asymmetric Clustering: - When the clustering method is registered, it may be K-mean normalization method which is selected for clustering with the normalization equation. The normalization equation when implemented with k-mean the cluster quality can be improved. The second method is of asymmetric clustering which is implemented to cluster the asymmetric data from the loaded dataset.

4. Apply mean shift and affinity metrics: - In this step, two operations are performed. In the first step mean shift algorithm is applied on the loaded dataset. In the mean shift algorithm, the mean value is calculated on the dataset and left shift operation is performed to simplify the operation of clustering. The second method is of affinity metrics, it is equation which is applied to find relationship between various elements of the dataset.

Copyright © 2019Author

ISSN: 0971-1260

Vol-22-Issue-3-July-2019

5. Apply MCL and S-clustering: - The MCL is the markov clustering algorithm, which is the unsupervised clustering graph based algorithm. This algorithm is fast and reliable and has good cluster quality. The main concept behind this algorithm is mathematical theory behind it, its position in cluster analysis and graph clustering, issues concerning scalability, implementation, and benchmarking, and performance criteria for graph clustering in general. The second method is S-clustering which is applied to cluster the data on the basis of graph methods

6. Plot and make clustering and normalize: - In the previous step, two methods are applied which are MCL and S-cluster, to cluster the data. In this method clustered data will be plotted. When the data is plotted, the method of normalization will be applied on the plotted data to improve the cluster quality.

Normalization and K-means working steps are:

Input: Data set $P = \{p1, p2, \ldots, pn\}$, cluster number k.

Step 1 Compute the distance matrix W, construct similarity matrix S according to W, where $W(i, j)$ is the distance between pi and pj, $i = 1, 2, \ldots, n$;

Step 2 Calculate the Laplacian matrix, $L = D - S$;

Step 3 Compute the first k eigen vectors $\{v1, \ldots, vk\}$ of the generalized eigen problem $Lv = \lambda Dv$;

Step 4 Let $V \in Rn \times k$ be a matrix composed of the vectors $\{v1, \ldots, vk\}$ as columns;

Step 5 For $i = 1, \ldots, n$, let $yi \in R1 \times k$ be the vector corresponding to the ith row of V;

Step 6 Cluster the points $\{yi \in R1 \times k \mid i = 1, 2, \ldots, n\}$ with the k-means algorithm into clusters $C1, \ldots, Ck$, if $yi \in Cj$ then $pi \in Pj$, $1 \le i \le n$, $1 \le j \le k$.

Output: k clusters $P1, \ldots, Pk$.

Copyright © 2019Author

ISSN: 0971-1260

Vol-22-Issue-3-July-2019

Affinity and Mean Shift Working Steps

Input: Data set $P = \{p1, p2, \ldots, pn\}$, > 0, $\delta > 0$, user-specified upper threshold $Cmax \geq 2$ for cluster number to be testified, user-specified maximum number of neighbors $Kmax \geq 2$.

Step 1 Calculate the distance matrix W;

Step 2 For $i = 1, 2, \ldots, n$, sort the ith row of W, then calculate $pi\ K$, which is the Kth neighbor of pi , $K = 2, \ldots, Kmax$;

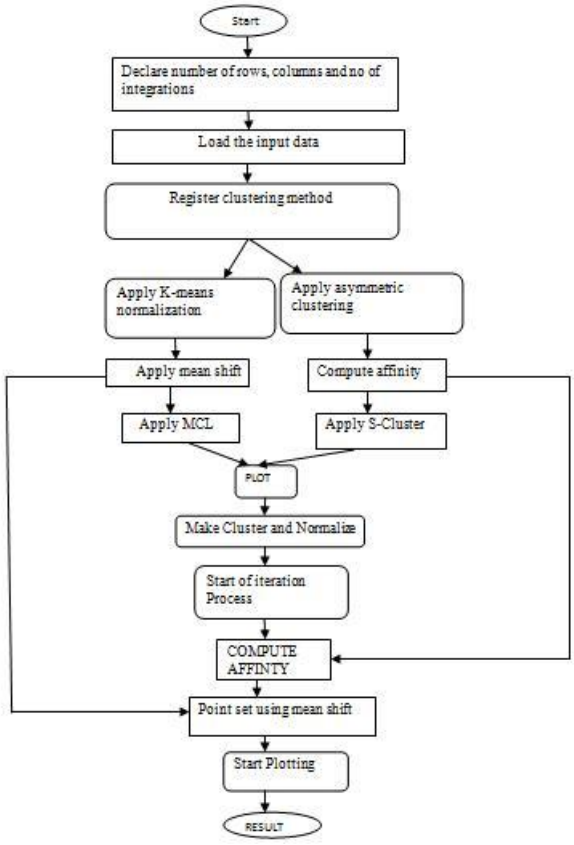
Step 3 For $K = 2, \ldots, Kmax$ run step 4~5;

Step 4 Calculate the similarity matrix S, where $S(i, j) = \exp(\ )$;

Step 5 For every $k = 2, \ldots, Cmax$, make use of the Meil˘a–Shi spectral clustering algorithm to cluster the data set P into k clusters and calculate the value of index Ratio(k) for obtained clusters;

Step 6 To determine whether the candidate cluster number $2 \leq k \leq Cmax$ is an -reasonable and δ-stable cluster number according to the results of step 4 and step 5;

Output: The set of reasonable and δ-stable cluster numbers.



Copyright © 2019Author

ISSN: 0971-1260

Vol-22-Issue-3-July-2019

Figure 4.1 Proposed Methodology's Flowchart

V. EXPERIMENTAL RESULT

In this paper, an asymmetric clustering has been implemented. After that a novel technique has been proposed to enhance asymmetric clustering to improve its noise, time and accuracy. The simulation has been done on MATLAB.

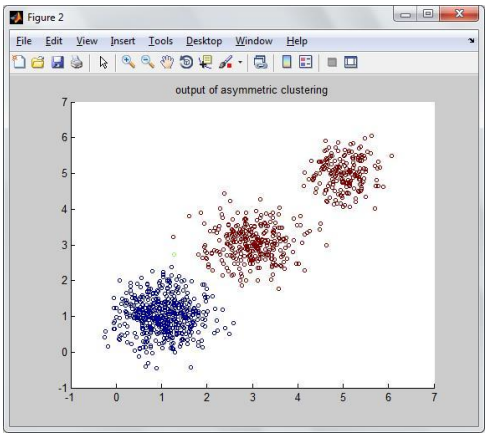
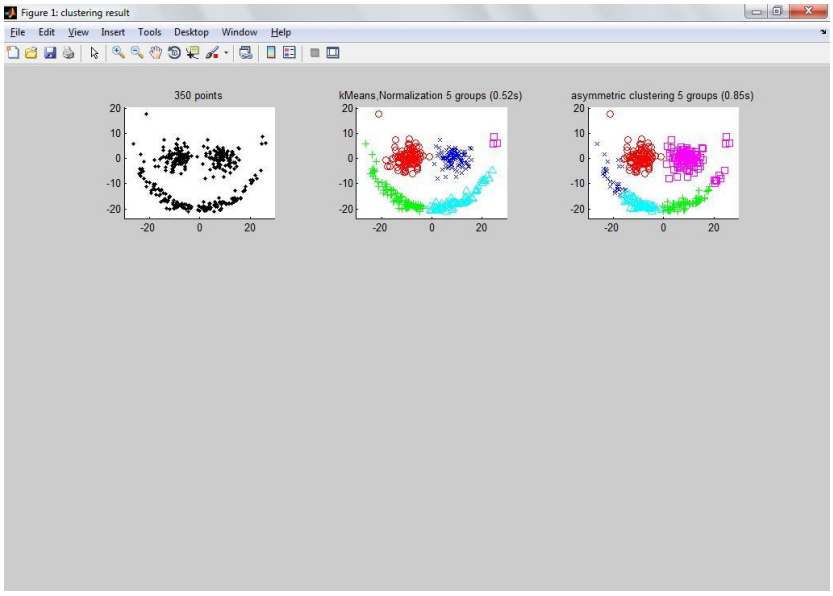


Figure 5.1 Output of asymmetric clustering

As illustrated within the figure 5.1, the dataset that is loaded have been scattered and scattered information is clustered uneven in keeping with uneven among the loaded statistics. Noise becomes more in this work due to scattering.



Copyright © 2019Author

ISSN: 0971-1260

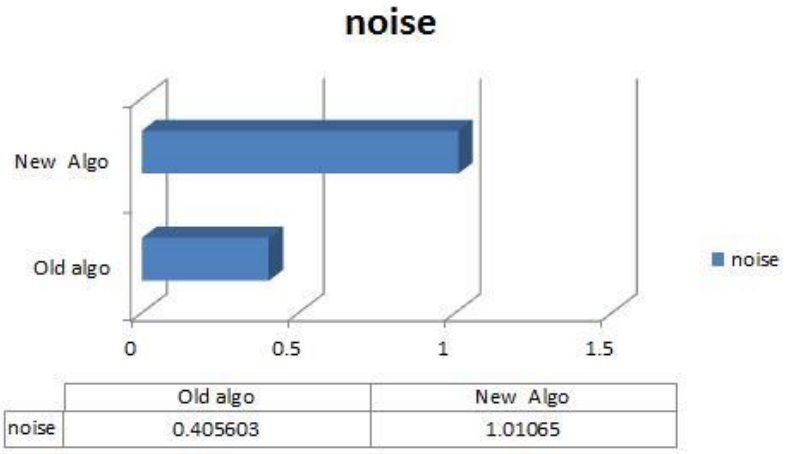
Vol-22-Issue-3-July-2019

Figure 5.2 Enhanced asymmetric clustering's output

The three different procedures as defined in this snapshot, as illustrated in figure 5.2. Data sets have been scattered at random in the first snapshot. The MCL and S-Clustering is applied for graphically displayed in the second figure. For data clustering, the asymmetric clustering is applied in the third figure.

VI.SUMMARY AND CONCLUSION

In the term of accuracy and performance, there are numerous algorithms that work well with simple data sets, but when these algorithms have to work with mixed and tightly coupled different data sets their performance in the term of accuracy is reduced. Asymmetric algorithms are enhanced in the previous asymmetric algorithm along with techniques such as k-means normalization, mean shift, Markov Cluster Algorithm, S-Cluster techniques to improve the efficiency of asymmetric clustering in terms of noise reduction and accuracy improvement. Noise is calculated approximately 0.405603(fig. 5.1) in the previous technique and time taken for execution is 1.00129 sec. The calculated noise is 1.01065 in the enhanced version of this asymmetric clustering algorithm and the time for execution is.09 sec. There is quite a big difference in both the preceding and enhanced asymmetric clustering accuracy. In previous asymmetric clustering accuracy is only 57.14 percent and 92.86 percent is seen in improved asymmetric clustering algorithm.



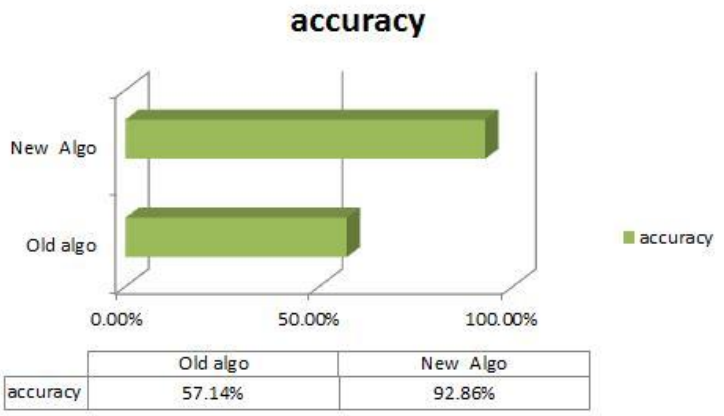
	Old algo	New Algo
noise	0.405603	1.01065

Figure 6.1 Comparison on basis of Noise

Copyright © 2019Author

ISSN: 0971-1260

Vol-22-Issue-3-July-2019



	Old algo	New Algo
accuracy	57.14%	92.86%

Figure 6.2 Accuracy of the two algorithms



	Old algo	New algo
Run time(In sec)	3.03	1.01

Figure 6.1Accuracy in terms of run time.

References

[1] Hao Huang, YunjunGao,KevinChiew Lei Chen Qinming He ” Towards Effective and Efficient Mining of Arbitrary Shaped Clusters” *Department of Computer Science and Engineering, Hong Kong University of Science and Technology, China,* ICDE Conference 2014

Copyright © 2019Author

[2] Satoshi Takumi and SadaakiMiyamoto,"*Top-down vs Bottom-up methods of Linkage for Asymmetric Agglomerative Hierarchical Clustering*", International Conference on granular Computing, 2013

[3] SumitGarg and Arvind K. Sharma, **"***Comparative Analysis of Data Mining Techniques on Educational Dataset",* International Journal of Computer Applications (0975 –8887) Volume 74–No.5 , July 2013

[4]] Gunnar Carlsson et.al , "Hierarchical Quasi-Clustering Methods for Asymmetric Networks",*Proceedings of the 31st International Conference on Machine Learning, Beijing, China, 2014. JMLR:W&CP volume 32, 2014*

[5] HuiXiong, GauravPandey, Michel and Vipun, *"Enhancing Data Analysis with Noise Removal",* IEEE Transaction on Knowledge and Data Engineering, Vol. 13, 2013

[6] Satoshi Takumi and SadaakiMiyamoto,"*Top-down vs Bottom-up methods of Linkage for Asymmetric Agglomerative Hierarchical Clustering*", 2012 International Conference on granular Computing

[7] Ertoz, L., Steinbach, M., and Kumar, V., *"Finding clusters of different sizes, shapes, and densitie dimensional data"*, In Proc. of SIAM DM'03.

[8] Ester, M., Kriegel, H.P., Sander,J., and Xu, X., *" A density-based algorithm for discovering clusters databases with noise",* in Proc. of 2nd Int. Conf. on Knowledge Discovery and Data Mining(KDD-96),AAAI Press, 1996, pp. 226-231.

[9] Fayyad, U. and Grinstein ,G., *"Information Visualization in Data Mining and Knowledge Discovery*", M 2001, pp. 182-190.

[10] Han, J., Kamber, M., and Tung, A. K. H., "Spatial clustering methods in (eds.), *Geographic Data Mining and Knowledge Discovery,*Taylorand Francis, 2001.

[11] Harel, D. and Koren, Y., *"Clustering spatial data using random walks",* In Proc. 7th and Data Mining(KDD-2001),ACM Press, New York, pp. 281-286

 Copyright © 2019Author

ISSN: 0971-1260

Vol-22-Issue-3-July-2019

[12]Satoshi Takumi and Sadaaki Miyamoto, "*Top-down vs Bottom-up methods of Linkage for Asymmetric Agglomerative Hierarchical Clustering*",International Conference on granular Computing, 2012

Copyright © 2019Author