ISSN 2348 - 2249
Volume-2, Issue-3, July-September, 2014
International Journal for Multi Disciplinary Engineering and Business Management
(IJMDEBM)

Hybrid of K-Means and Hierarchal Algorithms to Optimize Clustering

Jasvinder Kaur
Research Scholar, Department of Computer Engineering
Punjabi University, Patiala
India
kantu88@gmail.com

Gaurav Gupta
Department of Computer Engineering
Punjabi University, Patiala
India
gaurav_shakti@yahoo.com

Abstract - Data mining is the exploration and analysis of large quantities of data in order to discover valid, novel potentially useful and ultimately understandable patterns in data. This paper presents the concept of data mining and aims at providing combination of K-Means Clustering and Hierarchal Clustering to provide a hybrid clustering which is very useful in providing stability in clustering mechanism. This research has great significance as it provides efficient clustering of data sets and can also be applicable on text clustering. Make the information retrieval process for fast as compared to single clustering algorithm. Create more relevant clusters. Hybrid clusters make our research more accurate and Fast as compared to find the information from raw data.

Keywords – Data mining, Data mining techniques, Clustering, Clustering applications, K-means clustering, Hierarchical clustering, Netbeans IDE, Image Retrieval through clustering.

I. INTRODUCTION

Data mining (sometimes called data or knowledge discovery) is the process of analyzing data from different perspectives and summarizing it into useful information - information that can be used to increase revenue, cuts costs, or both. Data mining software is one of a number of analytical tools for analyzing data. It allows users to analyze data from many different dimensions or angles, categorize it, and summarize the relationships identified. Data Mining is core part of Knowledge Discovery Database (KDD). Many people treat Data mining as a synonym for KDD since it's a key part of KDD process. Knowledge discovery as a process is depicted in Figure 1 and consists of an iterative sequence of the following steps:

• Data Cleaning-To remove noise or irrelevant data.

• Data Integration - Where multiple data sources may be combined.

• Data Selection - Where data relevant to the analysis task are retrieved from the database.

• Data Transformation - Where data are transformed or consolidated into forms appropriate for mining by performing summary or aggregation operations.

• Data Mining - An essential process where intelligent methods are applied in order to extract data patterns.

• Pattern Evaluation -To identify the truly interesting patterns representing knowledge based on some interestingness measures.

• Knowledge Presentation- knowledge representation techniques are used to present the mined knowledge to the user.

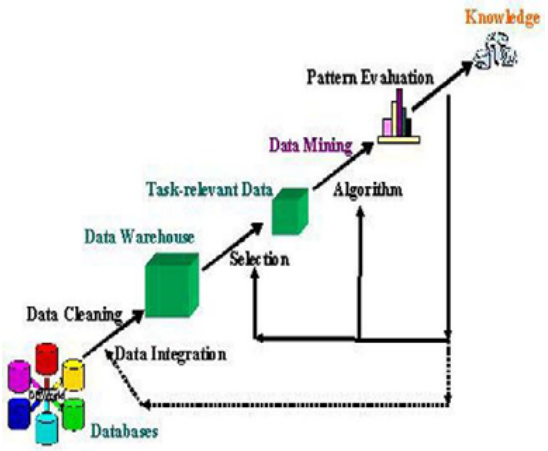


Fig.1 Preprocessing and Mining

II. DATA MINING TECHNIQUES:

There are several major data mining techniques have been developed and used in data mining:

A. Association
B. Classification
C. Clustering
D. Prediction
E. Sequential Patterns

III. CLUSTERING

Clustering is the process of partitioning a set of data (or objects) into a set of meaningful sub-classes, called clusters. It helps users to understand the natural grouping or structure in a dataset. A good clustering method will produce high quality clusters in which the intra-class (i.e., intra-clusters) similarity is high and the inter-class similarity is low. The quality of clustering result also depends on both the similarity measure used by the method and its implementation. The quality of a clustering method is also measured by its ability to discover some or the entire hidden pattern.

ISSN 2348 - 2249
Volume-2, Issue-3, July-September, 2014
International Journal for Multi Disciplinary Engineering and Business Management
(IJMDEBM)

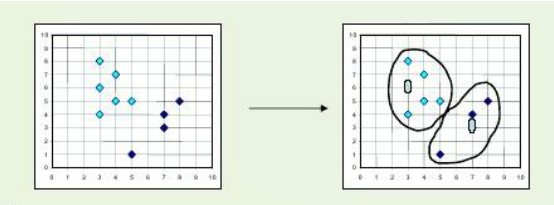


Fig.2 The result of Cluster analysis

Cluster analysis groups objects (observations, events) based on the information found in the data describing the objects or their relationships. The goal is that the objects in a group will be similar (or related) to one other and different from (or unrelated to) the objects in other groups. The greater the similarity (or homogeneity) within a group and the greater the difference between groups, the "better" or more distinct the clustering.

Example of clustering: In order to elaborate the concept a little bit, let us take the example of the library system. In a library books concerning to a large variety of topics are available. They are always kept in form of clusters. The books that have some kind of similarities among them are placed in one cluster. For example, books on the database are kept in one shelf and books on operating systems are kept in another cupboard, and so on. To further reduce the complexity, the books that cover same kind of topics are placed in same shelf. And then the shelf and the cupboards are labelled with the relative name. Now when a user wants a book of specific kind on specific topic, he or she would only have to go to that particular shelf and check for the book rather than checking in the entire library.

Applications

Clustering algorithms can be applied in many fields, for instance:

• Marketing: finding groups of customers with similar behaviour given a large database of customer data containing their properties and past buying records.

• Biology: classification of plants and animals given their features.

• Libraries: book ordering.

• Insurance: identifying groups of motor insurance policy holders with a high average claim cost; identifying frauds.

• City-planning: identifying groups of houses according to their house type, value and geographical location.

• Earthquake studies: clustering observed earthquake epicentres to identify dangerous zones.

• WWW: document classification; clustering weblog data to discover groups of similar access patterns

IV. K-MEAN CLUSTERING: THE ALGORITHM

The k-means algorithm (Lloyd, 1982) belongs to a family of algorithms known as optimization clustering algorithms. In this family of algorithms, clusters are formed such that some criterion of cluster goodness is optimized. That is, the examples are partitioned into clusters such that the clusters are optimal according to some measure. The name comes from the fact that k clusters are formed, where the centre of the cluster is the arithmetic mean of all vectors within that cluster.

$$J = \sum_{j=1}^{k} \sum_{i=1}^{n} \left\| x_i^{(j)} - c_j \right\|^2$$

Where $\left\| x_i^{(j)} - c_j \right\|^2$ is a chosen distance measure between a data point $x_i^{(j)}$ and the cluster centre , is an indicator of the distance of the n data points from their respective cluster centre.

The k-means algorithm is as follows:

1. Select k seed examples as initial centres (randomly generated vectors can also be used).
2. Calculate the distance from each cluster centre to each example.
3. Assign each example to the nearest cluster.
4. Calculate new cluster centres, where each new centre is the mean of all vectors in that cluster.
5. Repeat steps 2-4 until a stopping condition is reached

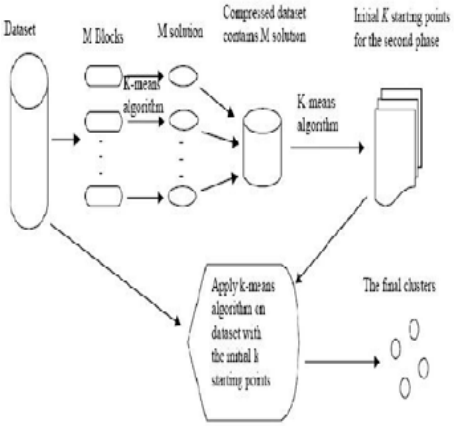


Fig.3 An overview of k-means

In the experiments reported here, the initial centres were vectors that were randomly selected from the data set, and the stopping criterion was based on the movement of the cluster centres: when vectors no longer changed clusters between iterations (the clusters had stabilized), the algorithm terminated. The number of clusters was set equal to the number of SOM output map neurons that were evaluated.

The disadvantage of k-means compared to SOM is that it does not perform vector quantization, that is, it does not naturally result in a form that can be easily visualized.

The advantage of k-means over SOM is that it is more computationally efficient and can thus run much faster.

V. HIERARCHICAL CLUSTERING

Hierarchical clustering algorithms divide into two categories: Agglomerative and Divisive Agglomerative clustering executes in a bottom–top fashion, which initially treats each data point as a singleton cluster and then successively merges clusters until all points have been merged into a single remaining cluster. Divisive clustering, on the other hand, initially treats all the data points in one cluster and then split them gradually until the desired number of clusters

ISSN 2348 - 2249
Volume-2, Issue-3, July-September, 2014
International Journal for Multi Disciplinary Engineering and Business Management
(IJMDEBM)

is obtained. To be specific, two major steps are in order. The first one is to choose a suitable cluster to split and the second one is to determine how to split the selected cluster into two new clusters. Many agglomerative clustering algorithms have been proposed, such as CURE, ROCK, CHAMELEON, BIRCH, single-link, complete-link, average-link, Leaders Sub leaders. One representative divisive clustering algorithm is the bisecting k-means method.

A hierarchical clustering creates a hierarchical decomposition of the given set of data objects. A hierarchical clustering can be classified as being either agglomerative or divisive, based on how the hierarchical decomposition if formed. The agglomerative approach, also called the bottom-up approach, starts with each object forming a separate group. It successively merges the objects that are close to one another, until all the groups are merging into one. The divisive approach, also called as top down approach, starts with all of the objects in the same clusters. In this, a cluster is split up into smaller clusters, until eventually each object is in one another. In this clustering, Dendrogram (Decompose data objects into a several levels of nested partitioning (tree of clusters), called a dendrogram are great for visualization. It provides hierarchical relations between clusters. it shown to be able to capture concentric clusters. In this clustering, it is not easy to define levels for clusters

.

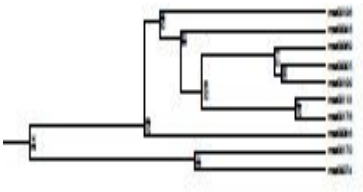


Figure 4: Dendogram

Agglomerative vs. Divisive.This aspect relates to algorithmic structure and operation. An agglomerative approach begins with each pattern in a Distinct (singleton) cluster, and successively merges clusters together until a stopping criterion is satisfied. A divisive method begins with all patterns in a single cluster and performs splitting until a stopping criterion is met.

Advantages

• It can produce an ordering of the objects, which may be informative for data display.

• Smaller clusters are generated, which may be helpful for discovery.

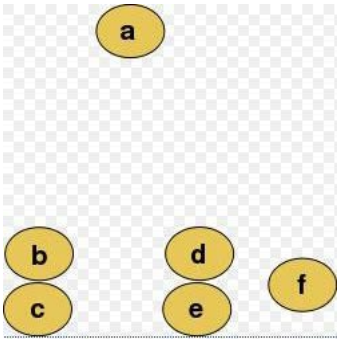


Fig.5 Example for Agglomerative Clustering

Objects that belong to a child cluster also belong to the parent cluster.

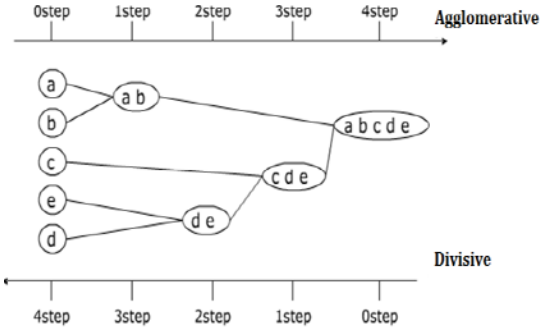


Fig. 7 combined diagram of Aglomerative&Divise

Algorithm Hierarchical Clustering

1. Construct one cluster for each document.
2. Join the t most similar clusters.
3. Repeat 2 until a stopping criterion is reached.

The result of agglomerative clustering is strongly dependent on the similarity measure. The single-link method defines the similarity between two clusters as the similarity between the two most similar objects, one from each cluster. This may result in elongated, locally similar clusters. For equally sized clusters (in volume), the complete-link method is a better choice. Here, similarity between two clusters is defined as the similarity between the two most dissimilar objects, one from each cluster. The time complexity of the agglomerative algorithms are O(n2) as they all need to compute the similarity between all objects to find the pair of objects that are most similar. Many agglomerative clustering algorithms have been proposed, such as CURE, ROCK, CHAMELEON, BIRCH, single-link, complete-link, average-link, and Leaders-Sub leaders.

Advantages

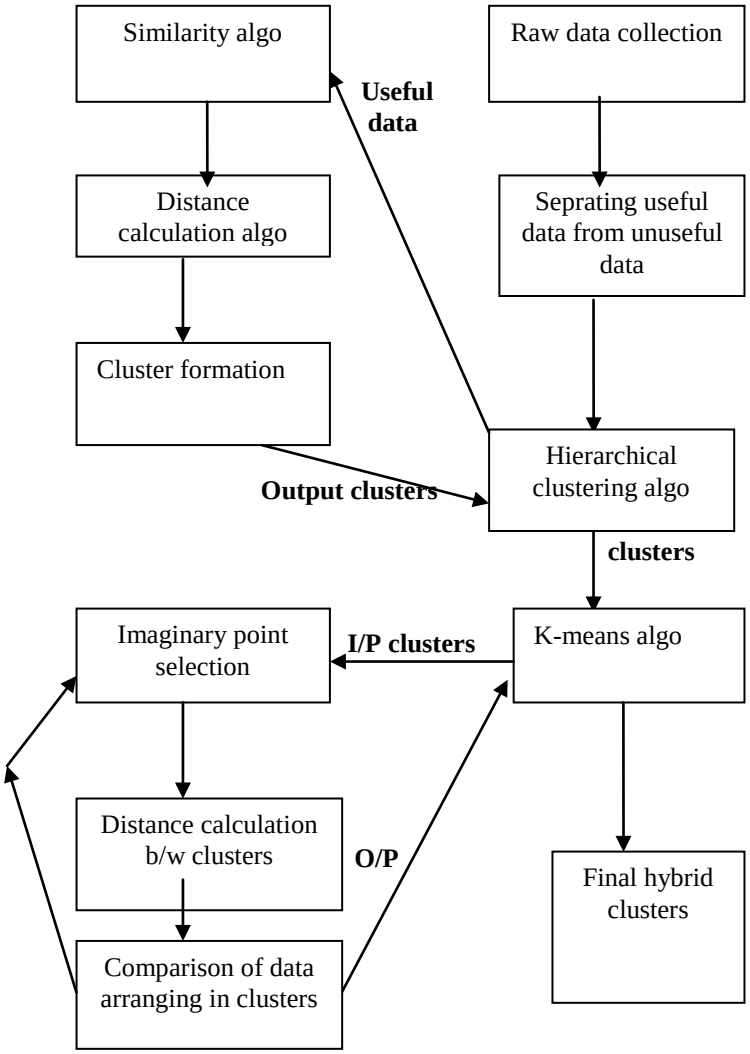
• It can produce an ordering of the objects, which may be informative for data display.

• Smaller clusters are generated, which may be helpful for discovery.

Flowchart of overall work

Flowchart of over project explains the overall working of project means to say how it works. By reading file it contains raw collection of data and separate useful data from un useful data after then similarity algorithm is applied on useful data after applying similarity algorithm distance calculation algorithm is applied on it this process results in the formation of clusters which comes as output clusters for hierarchical clustering algorithms .After that specify number of cluster user want user give input k=1,2,3…….so on. k-means algorithms performs Imaginary point selection on input clusters after that distance is calculated between clusters and comparison is found between data and arrangement of clusters are done Until final clusters are found according to user input or depends upon the value of k. After this final hybrid clusters are comes. This is the overall flowchart of Optimized Clustering Algorithm with Hybrid K-Means and Hierarchal Algorithms

ISSN 2348 - 2249
Volume-2, Issue-3, July-September, 2014
International Journal for Multi Disciplinary Engineering and Business Management
(IJMDEBM)



VI. RESULTS AND DISSCUSSION

Implementation

The dissertation work on Optimized Clustering Algorithm with Hybrid K-Means and Hierarchal Algorithms is implemented on java platform using net-beans. Java is a computer programming language that is concurrent, class-based, object-oriented, and specifically designed to have as few implementation dependencies as possible. It is intended to let application developers "write once, run anywhere" (WORA), meaning that code that runs on one platform does not need to be recompiled to run on another. Java applications are typically compiled to byte code (class file) that can run on any Java virtual machine (JVM) regard.

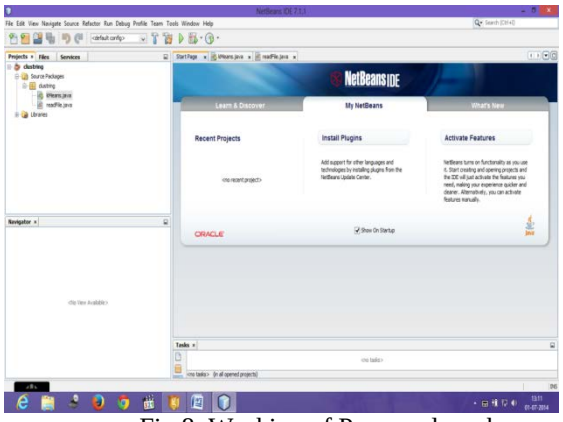


Fig.8 Working of Proposed work

Firstly, when we run the system it displays the following page that is shown in figure 8.This page shows the start page of NetBeansIDE after that we run file of clustering (k-means.java) under projects. The snap-shot of first page when we run our project shown in following figure 9.

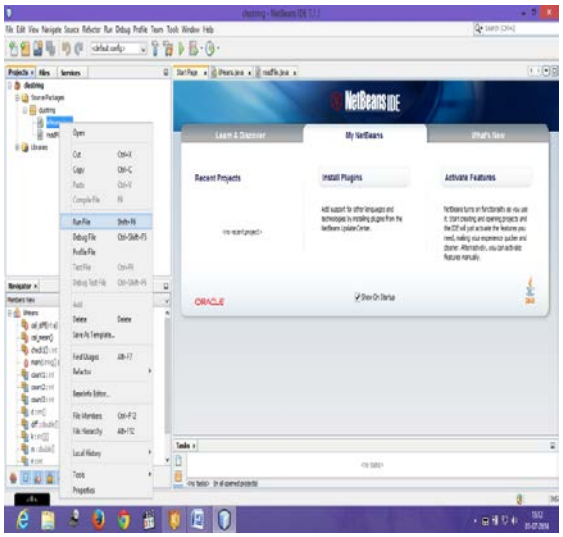


Fig.9 Run File of Clustering(k-mean.java) in Netbeans IDE

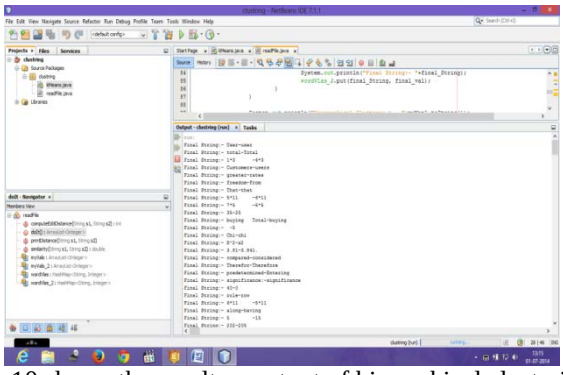


Fig.10 shows the result or output of hierarchical clustering

ISSN 2348 - 2249
Volume-2, Issue-3, July-September, 2014
International Journal for Multi Disciplinary Engineering and Business Management
(IJMDEBM)

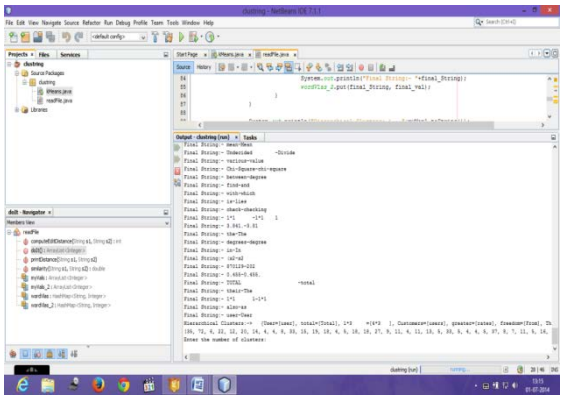


Fig.11 shows the hierarchical clustering in descending order

Hierarchical clustering is done in two orders ascending order or in descending order .In this project we perform descending order clustering. In descending order we have number of words or bunch of words in one cluster by comparing one by one word and check similarity and distance.

{[1][2][3][4][5]}One cluster contains many words.

{[1]}compare {[2][3][4][5]}

By conditions by checking similarity and distance.

{[ ]} {[ ]} {[ ]} {[ ]} {[ ]}

Output in the form of number of clusters.

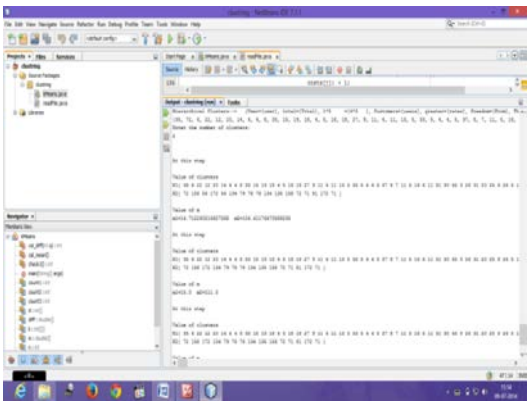


Fig.12 shows the output result by specifying number of clusters

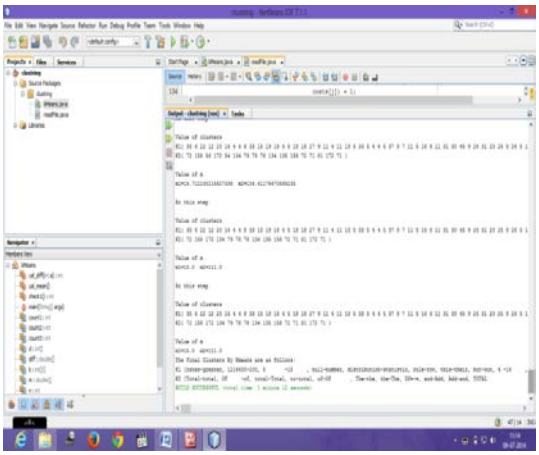


Fig.13 shows the final clusters and build successful total time

Advantages

1. Make the information retrieval process for fast as compared to single clustering algorithm.
2. Create more relevant clusters.
3.More meaningful information.
4. fast as compared to find the information from raw data.

VII. IMAGE RETRIEVAL THROUGH CLUSTERING OR CLUSTERING WORK IN ITS

Clustering is one of the most important parts of digital image analysis. Retrieval pattern-based learning with fast and refined clustering of the images is the most effective that aim to establish the relationship between the images with similar attributes. In related studies, various techniques have been proposed for retrieval of images based on component including color based and texture based fetching. In our research we have introduced a hybrid solution for image retrieval system with combination of K-Means and Hierarchical Image Clustering Algorithms which shows useful results for image fetching process. The main focus is to make image retrieval as fast as possible. For better image quality retrieval, we have implemented image quantization process for whole system so that we can access images stored in database easily.

Image matching is based on saturation and Color intensity for almost all images in database. Image retrieval system uses the Hybrid algorithm that is the combination of the K-Mean algorithm and the hierarchical algorithm. In our algorithm the k-mean clustering is performed first than we are using the hierarchical clustering. The image is given as the input to the propose system and then some segment of the image is taken and then that segment of the image is matched with the all the images in the database. These segment of the image taken is based on the color value RGB and these RGB value is compared with the images in the database we have taken the three segment based on the color value RGB and according to the three different segment of that hierarchy the image is matched from the database and the Percentage of the matched image is also shown for all the three section.

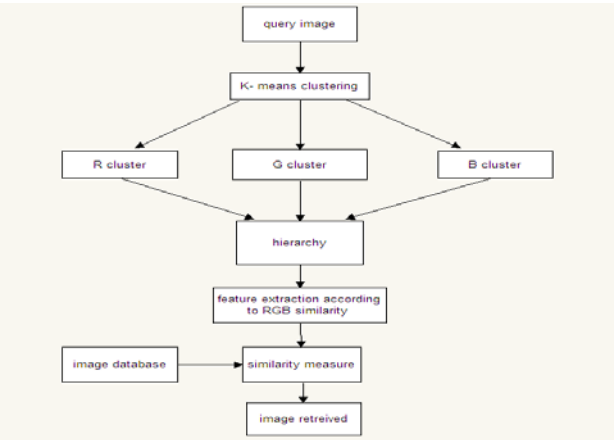


Figure.14 Flow Chart for proposed Image Retrieval System

ISSN 2348 - 2249
Volume-2, Issue-3, July-September, 2014
International Journal for Multi Disciplinary Engineering and Business Management
(IJMDEBM)

VIII. CONCLUSION FUTURE WORK

Our clustering algorithm is combination of two clustering algorithms which will make clusters with more efficiency and these clusters will be so meaningful and make our research more accurate and fast as compared to find the information from raw data that they will make the information retrieval faster.

The value of k in k-means algorithm can be determined by developing a new algorithm so that proper number of clusters according to data are generated. In future, we will improve the combined approach by integrating it to content image retrieval system so that combination can be used for larger applications**.**

REFRENCES

[1] Hong Yu, Xiaolei Huang, Xiaorong Hu, Hengwen Cai (2010) "A Comparative Study on Data Mining Algorithms for Individual Credit Risk Evaluation", International Conference on Management of e-Commerce and e-Government.

[2] Ji Dan, Qiu Jianlin (2010) "A Synthesized Data Mining Algorithm Based on Clustering and Decision Tree", 10th IEEE International Conference on Computer and Information Technology, CIT.

[3] Mohamed El far, Lahcen Moumoun, Mohamed Chahhou, Taoufiq Gadi (2010) "Comparing between data mining algorithms: "Close+, Apriori and CHARM" and "K-Means classification algorithm" and applying them on 3D object indexing", 10th IEEE International Conference on Computer and Information Technology, CIT.

[4] S.P.Latha (2007) "Algorithm for Efficient Data Mining", International Conference on Computational Intelligence and Multimedia Applications, Kavaraipettai.

[5] Wangjie Sun, Zhigao Zheng (2010) "An Advanced Design of Data Mining Algorithms", IEEE.

[6] Xiangyang Li, Nong Ye (2006) "A Supervised Clustering and Classification Algorithm for Mining Data with Mixed Variables", CIT.

[7] Zhu Can-Shi (2011) "A Study on the Application of Data Stream Clustering Mining through a Sliding and Damped Window to Intrusion Detection", Air Force Engineering University, China.

[8] M. Kuchaki Rafsanjani, Z. Asghari Varzaneh, N. Emami Chukanlo (2012) "A survey of hierarchical clustering algorithms", The Journal of Mathematics and Computer Science.

[9] A. S. Galathiya (2012) "Classification with an improved Decision Tree Algorithm", International Journal of Computer Applications.

[10] Navjot Kaur, Jaspreet Kaur Sahiwal, Navneet Kaur (2012) "Efficient K-Means Clustering Algorithm Using Ranking Method In Data Mining", International Journal of Advanced Research in Computer Engineering & Technology.

[11] Neelamadhab Padhy1, Dr. Pragnyaban Mishra 2and Rasmita Panigrahi3 "The Survey of Data Mining Applications And Feature Scope" International Journal of Computer Science, Engineering and Information Technology (IJCSEIT), Vol.2, No.3, June 2012.

[12] Y. Ramamohan, K. Vasantharao, C. Kalyana Chakravarti, A.S.K.Ratnam,`` A Study of Data Mining Tools in Knowledge Discovery Process", International Journal of Soft Computing and Engineering (IJSCE),Volume-2, Issue-3, July 2012.

[13] Monika Yadav, Mr. Pradeep Mittal, "Web Mining: An Introduction" International Journal of Advanced Research in Computer Science and Software Engineering, Volume 3, Issue 3, March 2013.

[14] Shailendra singh Raghuwanshi, PremNarayan Arya, "Comparsion of K-means and Modified K-mean algorithms for Large Data-set",International journal of Computing, Communications and networking", Volume 1, No.3, November- December 2012.

[15] Chunfei Zhang, Zhiyi Fang,"An Improved K-mean Clustering Algorithm" Journal of Information & Computational Science 10:1 (2013) 193-199.

[16] C. Lakshmi Devasena," A Hybrid Image Mining Technique using LIM-based Data Mining Algorithm", International Journal of Computer Applications, Vol. 25, No.2, July 2011, pp. 11-15

[17] X. Liu and P. He, "A Study on Text Clustering Algorithms Based on FrequentTermSets," in Advanced Data Mining and Applications, 2005, pp. 347-354.

[18] J. Han and M. Kamber, Data Mining: Concepts and Techniques, 2nd ed.: Morgan Kaufman Publishers, 2006.

[19] B. Fung, K. Wang, and M. Ester, "Hierarchical document clustering using frequent itemsets," in Proceedings of the SIAM International Conference on Data Mining, 2003.

