Journal of Advanced Zoology

ISSN: 0253-7214
*Volume **44** Issue **S-7 Year 2023** Page **338:346***

A Novel Information Extraction Approach Employing Hierarchical Clustering Techniques with A Focus on Attribute Similarity Values

D. Saravanan[1] , Topunuru Kaladhar[2]

[1]Faculty of Operations & IT ICFAI Business School (IBS), Hyderabad
[2]Department of Economics, ICFAI Faculty of Social Science The ICFAI Foundation for Higher Educatio (IFHE) (Deemed to be university u/s 3 of the UGC Act 1956) Hyderabad-India.

**Corresponding author's E-mail: D. Saravanan*

Article History

Received: 06 June 2023
Revised: 05 Sept 2023
Accepted: 13 Dec 2023

Abstract

Data mining technique developed in vast manner. It mainly used over in information retrieval, information filtering, Data retrieval techniques. In scientific world data mining techniques used over in hospital management, Schools, colleges, public libraries, shopping complexes, space research foundation. Data Mining refers to extracting the information. Many Algorithms are used over in Data mining area. Agglomerative hierarchical technique uses top down approach. Certain arrangements are made over in this algorithm in hierarchical manner. Hierarchical algorithm mainly used for time consuming process.

Keywords: *Data Exploration, Average Property Value, Document Comparison, Data Consolidation, Merging, Form One Group, Similarity Values*

1. Introduction

The agglomerative hierarchical clustering algorithm is effective in removing unwanted data from given input sets. When compared with other clustering techniques, this algorithm excels at eliminating unwanted or noisy data. It operates based on a tree structure, employing both top-to-bottom and bottom-to-top techniques. This method provides a valuable tool for researchers to understand where information is divided and where datasets are combined to create larger clusters. It aids researchers in splitting or combining entire sets, forming comprehensive cluster datasets. The algorithm functions by assessing the similarity between datasets. Data items that are closer are grouped in one dataset, while items with higher differences are placed in other datasets. In the tree structure, the algorithm calculates the link difference among data points. Small link distances are kept at the same level, while larger distances are placed at other levels. This method is beneficial for users who need to organize datasets in a hierarchical manner. In this work, time complexity and memory requirement are reduced. It outperforms the existing clustering algorithm. Hierarchical algorithm uses frequency values. Frequency values arrangement determine the particular persons identification. Similarity between the values are calculated. Using frequency values data is retrieved .Data can retrieved in frequent manner. Extracting data has clear ideas by using this technique. Accurate result will be produced over in this algorithm.

In today's fast-paced digital world, technology empowers both researchers and common users to generate vast amounts of data. This data is created and utilized by the user community without the need for knowledge from any specific source. It serves various purposes, including data segregation, identification, and extraction. Researchers must pay special attention to reconstructing or modifying this data due to its increasing production every year, influenced by various factors. Amidst this substantial dataset, extracting the required content becomes a challenging task for users. Even search engines do not yield consistent results when researchers submit identical queries. The results vary depending on the search engine, as depicted in the figure below. None of the search engines produce identical information, even when using the same key search terms .Users understand that, due to the proliferation of techniques, information production has increased. Consequently, users may not obtain the same information or content, even when using the same key terminologies. This phenomenon also applies to data. Given the high availability of data, users may not retrieve specific information in a

single search. Users often need to refine their searches, conducting one or more searches to obtain the desired content.

Clustering is performed based on the technique adopted. For example, in model-based clustering, information is grouped based on the user's chosen model. Data sets are grouped according to the technique or model selected by researchers. Unknown data sets are grouped based on the chosen model and fit into the model. In centroid clustering, for instance, data sets are initially scanned, and differences among the data are calculated. Eventually, the data will form clusters near the center. This process may involve one or more scans, with the data fitting into the final group. Users grasp the concept that data formation depends on the clustering methods chosen by researchers, resulting in different groupings for different data sets. The resultant cluster or group be subject to on the magnitude and complication of the information used. Clustering can sometimes be formulated in the first step itself or may take more than five or six steps, depending on the nature of the initial data sets. The letter 'R' indicates' repeating' or' replications,' representing the number of steps taken for initial or final cluster formations. Most cluster formations hinge on the space among binary information facts. If the distance is significant, the information facts are classified into different groups. Conversely, if the distance is small, they are kept in the same group. Researchers must initially calculate the distance among data sets to enhance cluster formation.

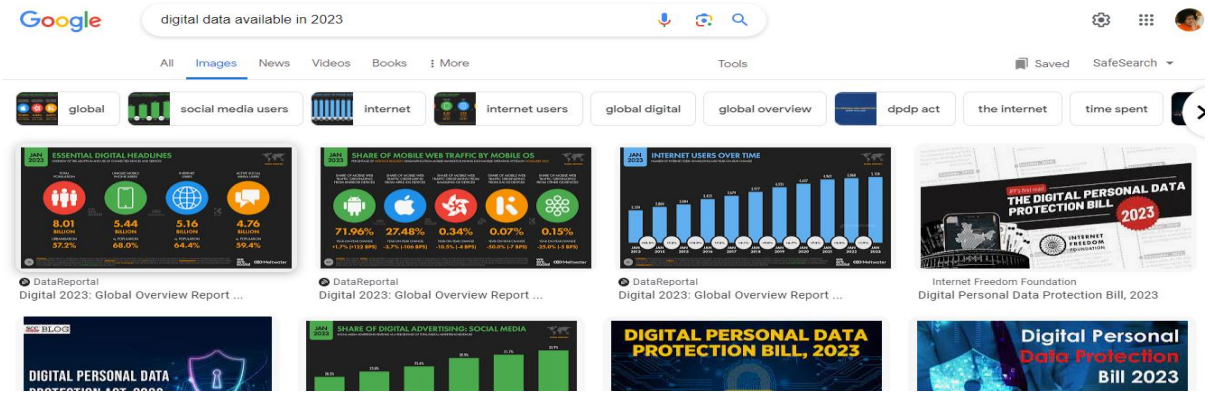

Fig 1: Digital data's availability using google search engine

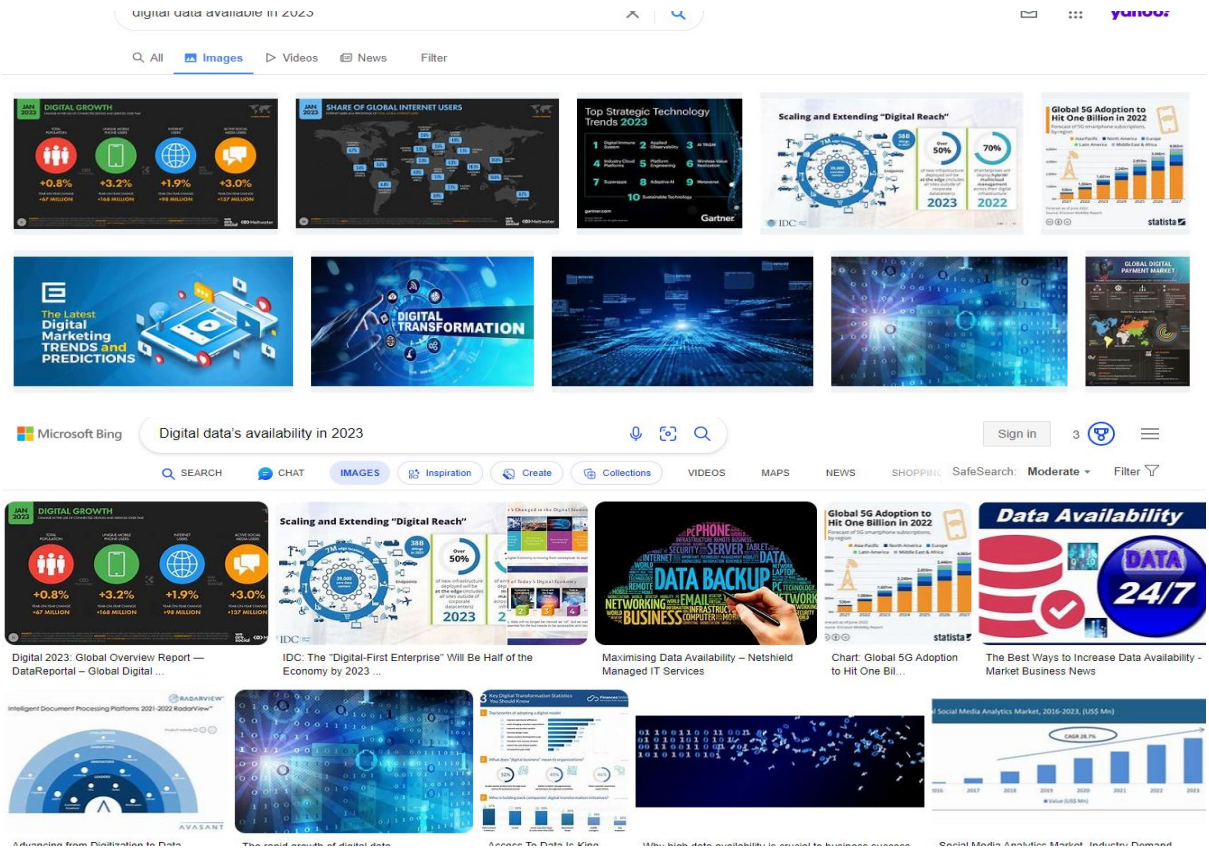

Fig 2: Digital data's availability using Bingo search engine.

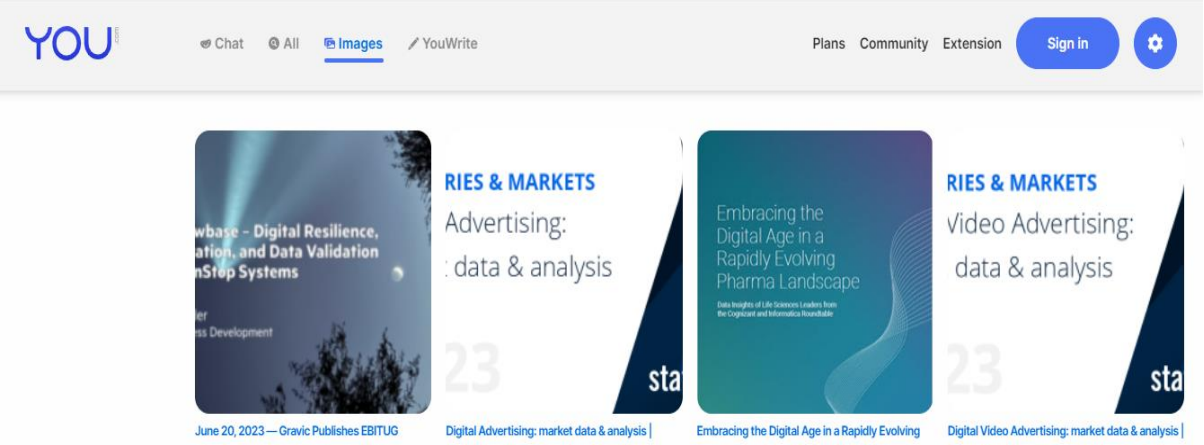

Fig 3: Digital data's availability using You search engine.

Problem Statement

The formation of clusters depends on the input datasets. The quantity of groups and outliers is firm founded on this information. Cluster formation is aligned with the researcher's objectives, specifying the type of data required for their operations. The formation can be adjusted based on the requirements, either increasing or decreasing clusters. In each scenario, the formation may increase or decrease, but it never reaches zero. This signifies the formation of groups based on the researcher's objectives to fulfil the specified goals. The researcher may obtain either a few clusters or a greater number, depending on the input set and research objectives. These groups are formed based on various objectives of cluster formation, such as the similarity between data points or groups. Similar techniques are considered for the formation of outliers. Depending on the input quality, cluster or outlier formations are executed. The researcher adjusts objectives based on the required sets of groups or clusters. If more groups are needed, the similarity index is modified, resulting in more groups. Similarity between objects or text is identified, and similarity metrics are considered for cluster formations. These measures vary depending on the type of data. For instance, if the input is an image, properties such as image colour, average frame value, distance between image pixels, average frame position, and more are considered for cluster formation. Users can choose any of these parameters as input, determining the cluster formation. Each property is treated separately, guiding cluster or outlier formations. Similarly, if the input is text or other documents, properties such as title, text repetition, nouns, verbs, and other functions are considered for cluster formation.

Structure design

The proposed scheme model is exposed in Figure 4. In this model, point datasets are considered. Initially, the input datasets are collected from different resources and stored in the database. From the database, the required content is extracted, and grades are assigned based on the relevance of the given datasets. This grading is calculated based on the identification of relevance, and grades are assigned accordingly. Point calculation is then performed by identifying the occurrence of information or text. The points are calculated based on the frequency of occurrence; a higher number of occurrences results in a higher number of points, contributing to a higher grade. This point calculation is facilitated through the use of a similarity index, where T1 and T2 are defined as the similarity index for text comparison.

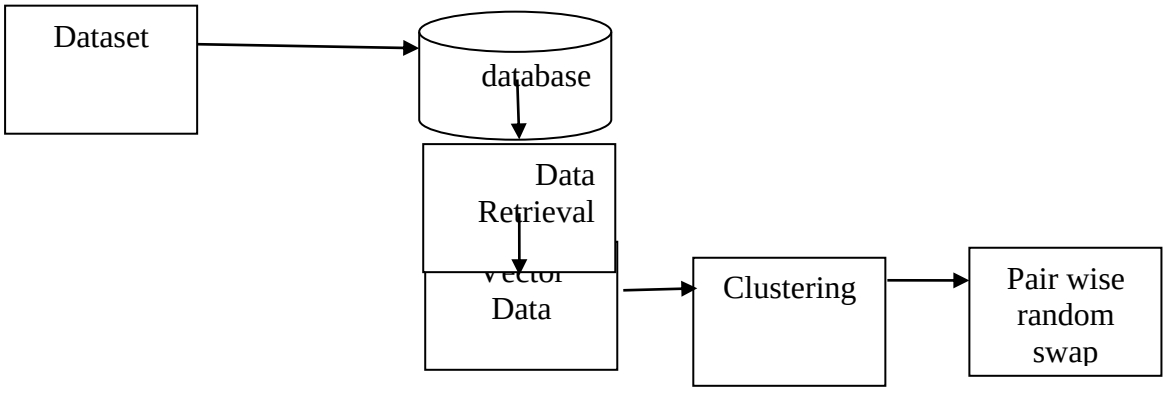

Fig 4: System Architecture Design for Clustering

Flow Diagram For Frequency Similarity

This flow diagram represents the flow of connection for hierarchical clustering algorithm. Frequency values are arranged in linear manner. Data is tabulated from data entry form. Frequency Threshold values are calculated.

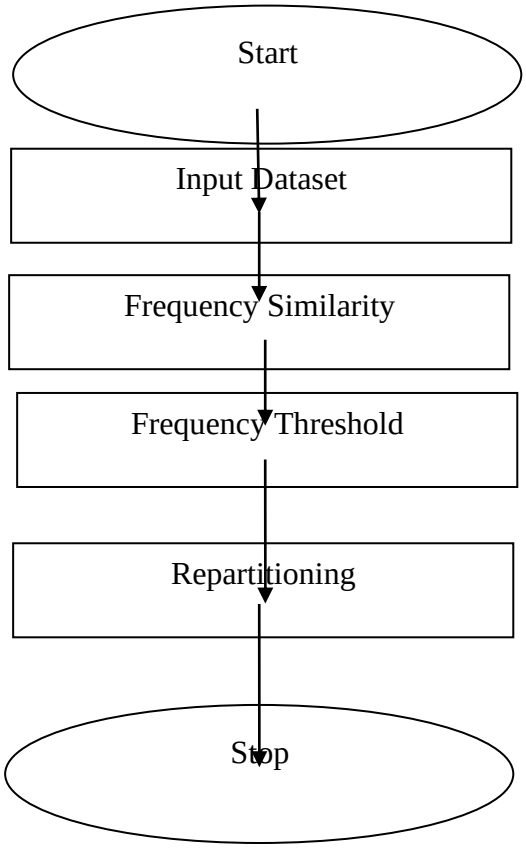

Fig 5: Flow diagram for frequency similarity

4. Experimental setup

4.1 Data Loading

4.2 Frequency Similarity

4.3 Initial Clustering

4.4 Repartitioning

4.1 Data Loading

In this process, the preprocessed dataset, which is stored in the database, is selected and converted into specific formats before being applied to the process for further operations.

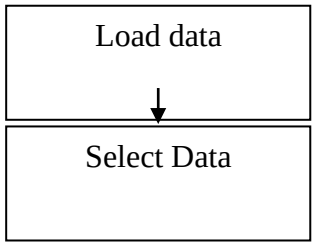

Fig 6: Data Selection Form

Frequency Similarity

After loading the data, the main operation is to identify the similarity among the datasets. This process helps us find the number of clusters or outlier formations. The identification is done through the consideration of various attributes in text or image data sets. The quantity of similar substances in the given datasets is crucial for determining similarity. Using attribute similarity, this process is

successfully carried out. Frequency values determine the location of the particular attribute. In the form values of the particular persons are arranged in hierarchical manner. Frequency Similarity is calculated by identifying the particular attribute. This categorization classified under 3 levels. Levels are monitored in each part. Frequency dissimilarity is also calculated. Computing the similarity between all pairs of matrices reveals the total number of attributes sharing the same properties. These values, or similarity properties, help researchers organize them into larger groups. Similar items are combined to form unique groups, while dissimilar items are placed into various separate groups. The only difference between the different cluster is processing with similarity data. In frequency similarity bottom up approach is also used. Values are retrieved from the database. In the data base certain locations are available.

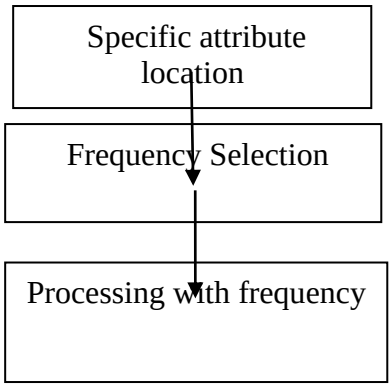

Fig 7: Frequency Similarity Method

Initial Clustering

In the initial clustering frequency based Similarity between entities are formed. To compute this value, users need to calculate the total attribute value available in the given document or frame. These values help researchers establish the acceptance threshold compared to each document set. If the values are higher than the average, they are considered and retained; if the values are lower than the average, the corresponding document or appearance is deemed a replica and abolished from the given datasets. Finally, all selected documents are together and deposited disjointedly for additional tasks. Hierarchical clustering is used for removing the outliers in the initial partitioning**.** In this process initialization of values took place. Unwanted information's are omitted in this process Garbage values are deleted. Initial clustering provided with certain parameter value. Valuess are retrieved in specific order .

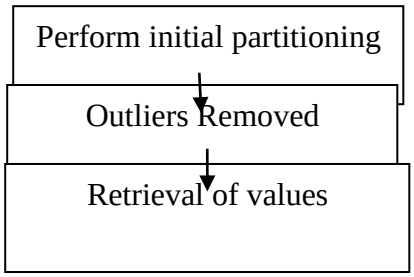

Fig 8: Initial Clustering Method

Repartitioning

Perform the repartitioning of preliminary grouping and yield the subgroups based on occurrence. Best fit clusters are processed. Repartition performed in top down manner. Outliers removed in the final clustering results. Hierarchical clustering for cluster formation is accomplished using two different techniques. In the first technique, entire cluster groups are divided into smaller clusters constructed on the relationship amongst the information arguments available within the groups. For instance, the entire population is divided into smaller groups, and each group is treated separately. In the second technique, the small groups are compared using item set properties, and the smaller groups with values very close to each other are grouped in the first stage. Similarly, the groups with values that are closer are merged to form a second, larger group. This process is repeated until all the small groups

are combined into a single group. In each technique, the choice between division or grouping techniques depends on the properties of the datasets. This requires any alteration limitation, except d, the amount of groups. In Repartitioning merging of clusters in tremendous manner. Various allocations are found over in this part. Best fit clusters played a main part. Outliers and various other garbage values are removed in this part. It mainly partition the attributes. And specifically it locate over in the particular location. It mainly retrieved from the graphical part.

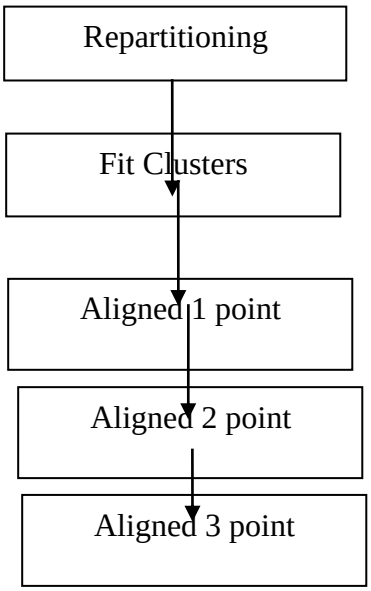

Fig 9: Repartitioning

Hierarchical Clustering Algorithms

Step 1:Initial Data Set Division:

- The initial data set, denoted as D, is divided into subsets {D1, D2, D3, D4,..., and so on), each considered as individual data set points.

Step2: Data Point Segmentation:

- The entire data set is further divided into smaller data points, each treated and tested separately.

Step 3: Attribute Similarity Comparison:

- Attribute similarities are utilized to compare each information topic with the average values of the information facts. Based on these values, items with higher similarities are grouped together, while values with lower similarities are treated as duplicates and removed.

Step 4: Distance Calculation:

- Distances among data points are calculated. For any given data points A and B, the dataset {A}{B} is determined as the minimum value{A}{B}, and this minimum is retained within the same group.

Step 5: Iterative Distance Calculation:

- Step 4 is repeated for all data sets until similar values are grouped together.

Step 6: Group Update and Similarity Measurement:

- Groups are updated by combining values that are closer. For instance, if C is a grouping variable, it is incremented (C = C + 1), and {A} and {B} are combined into a solo collection to form the subsequent aligned.

Step 7: Continued Iteration:

- This procedure is continual till all information arguments are consolidated into one comprehensive group.

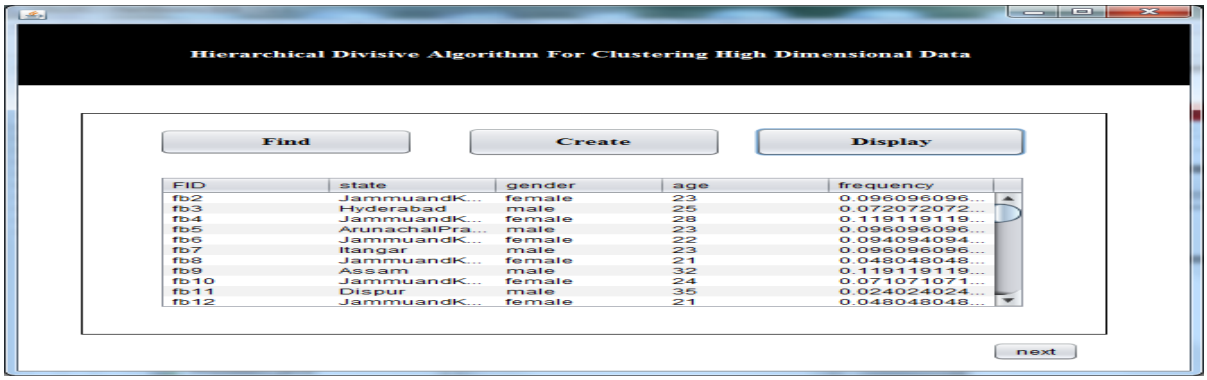

Fig 8: Persons Record form

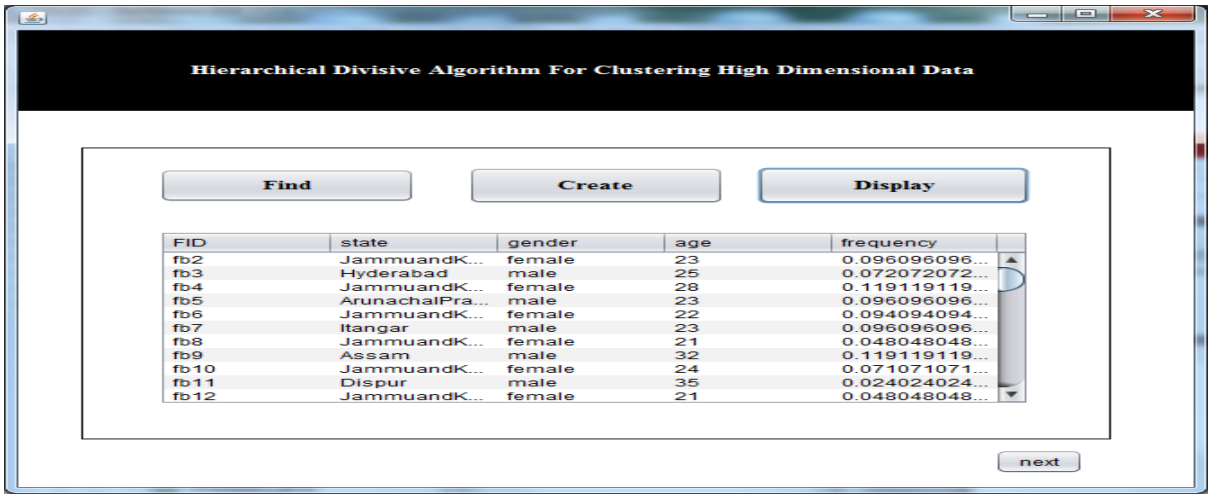

Fig 9: Frequency value display form

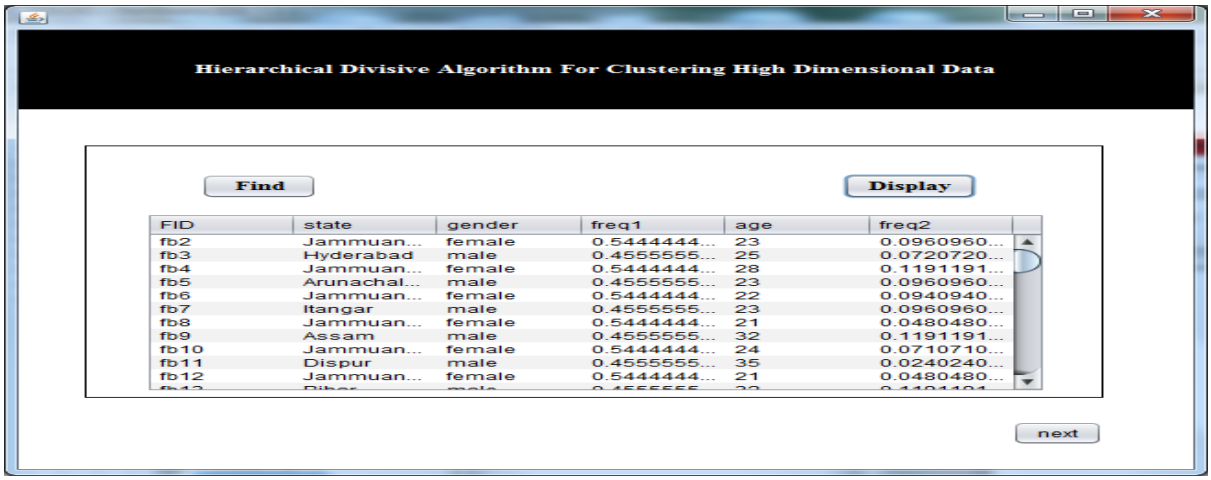

Fig 10: Frequency 2 Display form

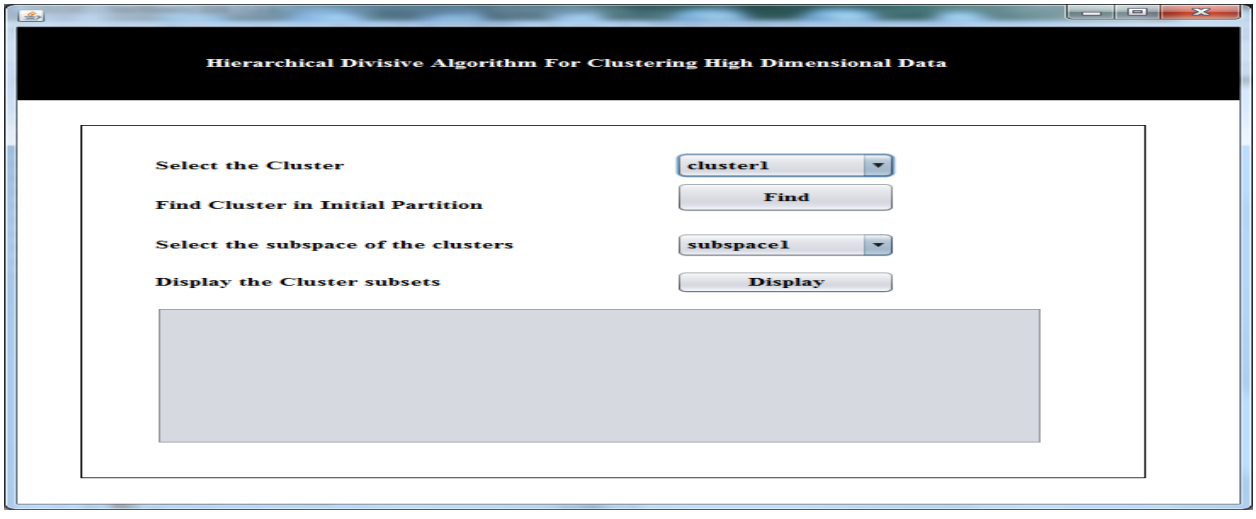

Fig 11: Cluster Selection form

4. Conclusion

This algorithm performs more efficient than any other algorithm. Records can be retrieved in easier manner. Less Manual implementation is required. Low computations are required. Performance requirements are low. Memory requirements are more in this process. Highly skilled engineers are required. High knowledge people can process this in efficient manner. Agglomerative Hierarchical clustering merge the data from smaller order, after merging of the data identification and detection of the attributes are easier. Mainly Level 1, Level 2, Level 3 is viewed for classification. In these levels age, gender frequency calculations are involved .Algorithm technique process this large data sets within a limited period of time. Storage facilities are available over here. Thus the algorithm produces more accurate result. Thus algorithm can be implemented in successful manner. New algorithmic techniques will be initiated. Time consuming must reduced less than present algorithm. Large number of records can be retrieved within a fraction of seconds using new implementation. In proposed system hierarchical levels are used for identification of the attributes. Advanced techniques will be involved in future work. Processing of the data is implemented limited techniques only. But the processing speed will be tremendous. New algorithm technique such as hierarchical divisible algorithm is used. By using this we can develop the future work.

References:

[1] Hierarchical Clustering of Large-scale Short Conversations Based on Domain Ontology 2008 international symposium on computer science and computational technology

[2] Interactive Visualization and Analysis of Hierarchical Neural Projections for Data Mining IEEE TRANSACTIONS ON NEURAL NETWORKS, VOL. 11, NO. 3, MAY 2000 615

[3] An Agglomerative Clustering Method for Large Data Sets International Journal of Computer Applications (0975 – 8887) Volume 92 – No.14, April 2014

[4] A hierarchical approach to represent relational data applied to clustering tasks proceedings of international joint conference san jose ,California USA,JULY 31-AUG 5 2011

[5] K. Jain, “Data clustering: 50 years beyond K-means,” Pattern Recognit. Lett., vol. 31, no. 8, pp. 651–666, 2010

[6] Segmentation of diffusion-weighted brain images using expectation maximization algorithm initialized by hierarchical clustering vol 102-No.17 fycal ramdani

[7] D.Saravanan,” Image frame mining Using indexing technique” Data Engineering and Intelligent Computing , SPRINGER Book series, Chapter 12 , Pages 127-137, ISBN:978-981-10-3223-3,July 2017.

[8] Xie, S-F. Chang, A. Divakaran and H. Sun, “Unsupervised Mining of Statistical Temporal Structures in Video Video Mining, eds. A. Rosenfeld, D. Doermann and D. DeMenthon, Kluwer Academic Publishers, 2003.

[9] I. Bartolini, P. Ciacci, and F. Waas, “Feedbackbypass: A New Approach to Interactive Similarity Query Processing,” Proc. 27th Int’l Conf. Very Large Data Base (VLDB ’01), pp. 201-210, 2001.

[10] R. Brunelli and O. Mich, “Image Retrieval by Examples,” IEEE Trans. Multimedia, vol. 2, no. 3, pp. 164-171, 2000.

[11] D.Saravanan” Effective Video Data Retrieval Using Image Key frame selection”, Advances in Intelligent Systems and computing , Pages 145-155,jan-2017.

[12]Fan, J.. Luo, H. "'Emantic Video Classification by Integrating Flexible MixtLre Model witlh Adaptive EM Algorithm",ACMSIGMM, 2003, pp.9-16.

[13]J. Zhang, W. Hsu and M. L. Lee. An Informationdriven Framework for Image Mining, in Proceedings of 12th International Conference on Database and Expert Systems Applications (DEXA), Munich, Germany, September 2001.

[14] Zhang Qi-dong, Wu Jian-hua, Gao Li-qun. Medical image retrieval based on non subsampled Contourlet transform and Zernike moments [J]. Chinese Journal of Scientific Instrument, 2009, 30(6): 1275-1280.

[15] Y.Liu .,Z.Li H.Xiong,X.Gao and J.wu,”Understand internal clustering validation measure”Proc 10th ICDNSW,Australia,2919,pp,911

[16] C.Agarwal,G,Hinnenburg and J.Jain”On the surprising behavior of distance measures in high dimensional space “in proc 8th ICDT, vol 1973,London,U.K.,2001,pp,877—886

[17] Xiong,J.wu and J.Chen,”K-means clustering versus validation measures:Adata distribution perspective”,IEEE Trans Syst.,Man,Cybern,B,cybern,vol, 39,no,2 pp,318-331,Apr 2009

[18] Guha,sudipto,Misra 2000 ROCK:clustering data streams IEEE transactions on clustering and data engineering.

[19]S.Z selim and M.A Ismail,K-means type algorithms:A generalized convergence theorm and characterisation of local optimity,IEEE transactions on pattern analysis and machine intelligence.

[20] jianfu Li,, jians huang LI,Huaiqing HE:Asimple and accurate approach to hierarchical clustering,journal of computational information system 7:7(2011)2577-2584.