Clustering Approach Recommendation System Using Agglomerative Algorithm

Mr.Anup D. Sonawane[1], Mr.Vijay Birchha[2]

[1] Mr. Anup D. Sonawane, PG Scholar, Computer Science and Engineering, SVCE, Indore, M.P., India
[2]Mr. Vijay Birchha, Asst. Professor, Computer Science and Engineering, SVCE, Indore, M.P., India

---***---

Abstract - *Clustering is one of the most imperative techniques and mostly used nowadays. Clustering applications are used extensively in various arenas such as artificial intelligence, pattern recognition, economics, ecology, psychiatry and marketing. There are several algorithms and methods have been developed for clustering problem. But problem are every time arises for discovery a new algorithm and process for mining knowledge for refining accuracy and productivity. There are several another issue are also exits like cluster analysis can contribute in compression of the information included in data. Clustering can be used to separator records set into a number of "motivating" clusters. Then, instead of processing the data set as an entity, we adopt the representatives of the defined clusters in our process. Thus, data compression is achieved. In this paper we projected a new and more effective approach based on agglomerative clustering. The proposed approach use simple calculation to identify based clustering approach.*

Key Words: **Cluster, Partition, hierarchical, accuracy, efficiency, agglomerative.**

1. INTRODUCTION

Clustering is unsupervised information because it doesn't method that predefined group linked with data objects. Clustering algorithms are engineered to find structure in the current data, not to categories future data. A clustering algorithm attempts to find natural groups of components (or data) based on some similarity.

A Cluster is a set of entities which are alike, and objects from different clusters are not alike. A cluster is an aggregation of points in the space such that the distance between two points in the cluster is less than the distance between any point in the cluster and any point not in it [1, 2].

All clusters are compared with respect to certain properties: density, variance, dimension, shape, and separation. The cluster should be a tight and compact high-density region of data points when compared to the other areas of space. From smallness and tightness, it paths that the step of dispersion (variance) of the cluster is small. The shape of the cluster is not known a priori. It is determined by the used algorithm and clustering criteria. Separation defines the degree of possible cluster overlap and the distance to each other [3].

2. CLUSTER PROPERTIES

Defining the properties of a cluster is a difficult task, although different authors emphasize on different characteristics. Boundaries of a cluster are not exact. Clusters vary in size, depth and breadth. Some clusters consist of small and some of medium and some of large in size. The depth refers to the range related by vertically relationships. Furthermore, a cluster is characterized by its breadth as well. The breath is defined by the range related by horizontally relationships [1, 4].

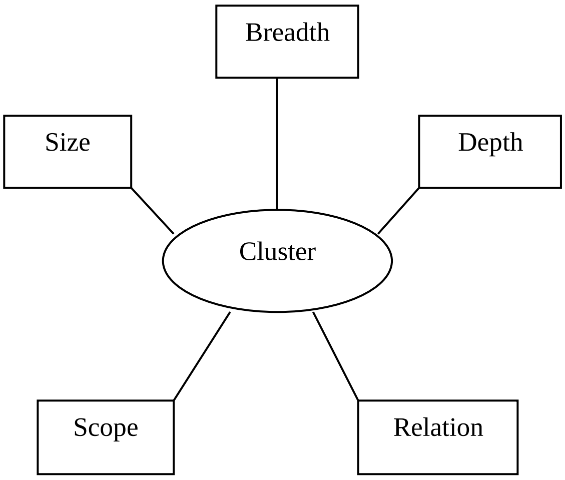

Figure 1 Properties of cluster

3. CLUSTER PROCESS

Cluster analysis is a useful technique for classifying similar groups of objects called clusters; objects in an exact cluster share many characteristics, but are very dissimilar to objects not belonging to that cluster. After having decided on the clustering variables we need to decide on the clustering procedure to form our groups of objects. This stage is critical for the analysis, as different processes require different results prior to analysis. These approaches are: hierarchical methods, partitioning methods and two-step clustering. Each of these procedures follows a different approach to grouping the most similar objects into a cluster and to determining each object's cluster membership. In other arguments, whereas an object in a confident cluster should be as similar as possible to all the other objects in the same cluster, it should likewise be as distinct as possible from objects in different clusters. An important problem in the application of cluster analysis is the decision regarding how many clusters should be derived from the data

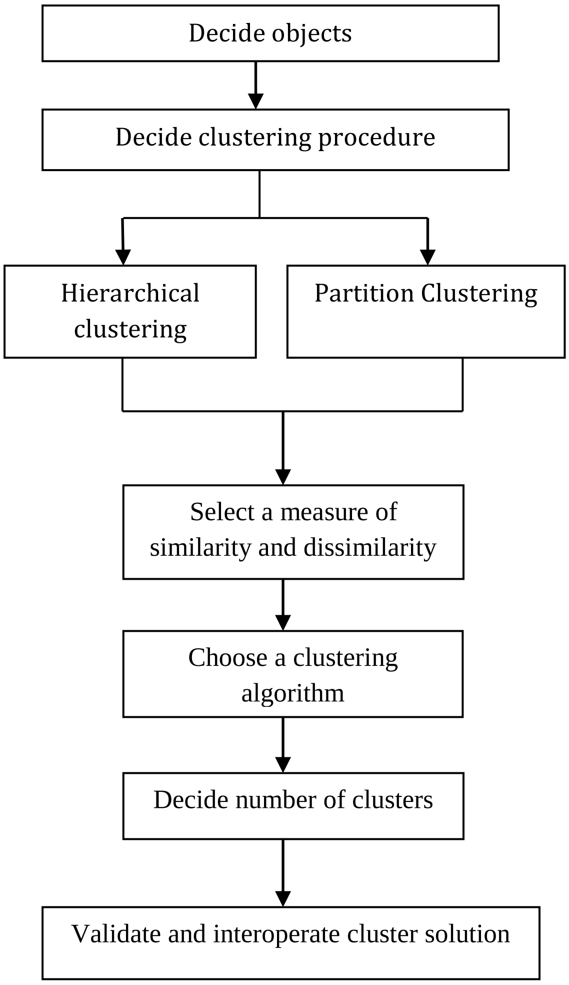
Decide objects
Decide clustering procedure
Hierarchical clustering
Partition Clustering
Select a measure of similarity and dissimilarity
Choose a clustering algorithm
Decide number of clusters
Validate and interoperate cluster solution

Figure 2 Clustering process

4. LITERATURE REVIEW

In 2010 Revati Raman et al proposed "Fuzzy Clustering Technique for Numerical and Categorical dataset". They presented a modified description of cluster center to overcome the numeric data only limitation of Fuzzy c-mean algorithm and provide a better characterization of clusters. The fuzzy k-modes algorithm for clustering unconditional records. They proposed a new cost function and distance measure based on co-occurrence of values. The measures also take into account the significance of an attribute towards the clustering process. Fuzzy k-modes algorithm for clustering unconditional records is extended by representing the clusters of categorical data with fuzzy centroids. The effectiveness of the new fuzzy k-modes algorithm is better than those of the other existing k-modes algorithms [5]

In 2011 Hussain Abu-Dalbouh et al proposed "Bidirectional Agglomerative Hierarchical Clustering using AVL Tree Algorithm". Proposed Bidirectional agglomerative hierarchical clustering to create a hierarchy bottom-up, by iteratively merging the closest pair of data-items into one cluster. The result is a rooted AVL tree. The n leafs resemble to input data-items (singleton clusters) needs to n/2 or n/2+1 steps to merge into one cluster, correspond to groupings of items in coarser granularities climbing towards the root. As observed from the time complexity and number of steps need to cluster all data points into one cluster perspective, the performance of the bidirectional agglomerative algorithm using AVL tree is better than the current agglomerative algorithms [6].

In 2011 Piyush Rai projected "Data Clustering: K-means and Hierarchical Grouping of cluster". They projected a proportional study. They display that flat clustering produces a single partitioning Hierarchical Clustering can give different partitioning depending on the level-of-resolution. Flat clustering needs the number of clusters to be specified Hierarchical clustering doesn't need the number of clusters to be specified Flat clustering is usually more efficient run-time wise Hierarchical clustering can be slow (has to make several merge/split decisions) No clear consensus on which of the two produces better clustering[11].

In 2011 Akshay Krishnamurthy et al "Efficient Active Algorithms for Hierarchical Grouping of cluster" They projected a general structure for active hierarchical clustering that repeatedly runs an off-the-shelf clustering algorithm on small subsets of the data and comes with guarantees on performance, measurement complexity and runtime complexity. They represent structure with a simple spectral clustering procedure and provide concrete results on its performance, showing that, under some assumptions [15].

In 2012 Dan Wei, Qingshan Jiang et al. projected "A novel hierarchical clustering procedure for gene Orders" .The proposed technique is evaluated by clustering functionally related gene sequences and by phylogenetic analysis. They presented a novel approach for DNA sequence clustering, mBKM, based on a new sequence similarity measure, DMk, which is extracted from DNA sequences based on the position and composition of oligonucleotide pattern. Proposed method may be extended for protein sequence analysis and Meta genomics of identifying source organisms of Meta genomic data [7].

In 2012 Neepa Shah et al Document Clustering: A Detailed Review" .They gave an overview of various document clustering methods, starting from basic traditional methods to fuzzy based, genetic, co-clustering, heuristic oriented etc. They also include the document clustering procedure with feature selection process, applications, challenges in document clustering, similarity measures and evaluation of document clustering algorithm is explained[12].

In 2013 Elio Masciari et al. proposed "A New, Fast and Accurate Algorithm for Hierarchical Clustering on Euclidean Distances "A simple hierarchical clustering algorithm called CLUBS (for Clustering Using Binary Splitting) is proposed in this paper. CLUBS is faster and more accurate than existing algorithms, including k-means and its recently proposed refinements. The procedure contains of a divisive stage and an agglomerative stage; during these two phases, the samples are repartitioned using a least quadratic distance criterion possessing unique analytical properties that. CLUBS derives good clusters without requiring input from users, and it is robust and impervious to noise, while providing better speed and accuracy than methods, such as BIRCH, that are endowed [8].

In 2014 J Anuradha, B K Tripathy projected "Attribute Dependency for Attention Deficit Hyperactive Condition". They projected a hierarchical clustering procedure to partition the dataset based on attribute dependency (HCAD). HCAD forms clusters of data based on the high dependent attributes and their equivalence relation. Proposed approach is capable of handling large volumes of data with reasonably faster clustering than most of the existing algorithms. It can work on both labeled and unlabeled data sets. Experimental results reveal that this algorithm has higher accuracy in comparison to other algorithms. HCAD achieves 97% of cluster purity in diagnosing ADHD [9].

In 2015 Z. Abdullah et al planned "Hierarchical Clustering Procedures in Data Mining" The proposed technique builds the solution by initially assigning each point to its own cluster and then repeatedly selecting and merging pairs of clusters, to obtain a single all inclusive clusters. The key factor in agglomerative procedures is the technique used to determine the pair of clusters to be merged at each step. Experimental results obtained on synthetic and real datasets demonstrate the effectiveness of the proposed various width cluster method [10].

In 2016 Amit Kumar Kar et al projected "Comparative Study & Performance Calculation of Different Clustering Methods in Data Mining". They evaluates the four major clustering algorithms namely: Partitioning methods, Hierarchical methods, Grid-based methods and Density-based methods and matching the performance of these algorithms on the basis of correctly class wise cluster building ability of algorithm[13].

In 2017 Shubhangi Pandit et al "An Enhanced Hierarchical Clustering Using Fuzzy C-Means Clustering Method for Document Content Analysis". They present effort a clustering method and projected using fuzzy c-means clustering algorithm for recognizing the text pattern from the huge data base. The projected work is also committed to advance the method of clustering for computing the hierarchical relationship among different data objects [14].

5. PROBLEM STATEMENT

The significant difficulties with ensemble based cluster study that these works have identified are as follows:

Distance measure: For numerical attributes, distance measures can be used. But identification of measure for categorical attributes in strength association is difficult.

Number of clusters: Finding the number of clusters & its proximity value is a hard task if the number of class labels is not known in advance. A careful analysis of inter & intra cluster proximity through number of clusters is necessary to create correct results.

Types of attributes: The databases may not necessarily cover characteristically numerical or categorical attributes. They may also contain other kinds like nominal, ordinal, binary etc. So these attributes have to be converted to categorical type to make calculations simple.

6. PROPOSED APPROACH

The projected method is used on generation of ensembles based cluster on the basis of few operations like mapping & combination. These operations can be performed with the help of two operators' similarity association & probability for correct classification or classifier analysis of cluster. In this planned approach our core aim is to classify the cluster partitioned data for hierarchical clustering. It may be represented via parametric representation of nested clustering.

1. Assign each object as individual cluster like c_1, c_2, c_3, .. c_n where n is the no. of objects
2. Find the distance matrix D, using any similarity measure
3. Find the nearby pair of clusters in the present clustering, say pair (r), (s), according to d(r, s) = mind (i, j) {i, is an object in cluster r and j in cluster s}
4. Merge clusters (r) and (s) into a single cluster to form a merged cluster. Store merged objects with its corresponding distance in tress distance Matrix.
5. Revalue distance matrix, D, by deleting the rows and columns corresponding to clusters (r) and (s). Adding a new row and column corresponding to the merged cluster(r, s) and old luster (k) is defined in this way: d [(k), (r, s)] = min d [(k), (r)], d [(k), (s)]. For further rows and columns copy the corresponding data from present distance matrix.
6. If all objects are in one cluster, stop. Otherwise, go to step 3.

7. Find association relation coefficient value with single, complete and average linkage methods

Table 1 Objects with coordinate value

Object	X	Y
A	4	4
B	8	4
C	15	8
D	24	4
E	24	12

Euclidean distance between p and q are denoted as

$$d(p,q) = \sqrt{(\mathbf{x1} - \mathbf{x2})^2 + (\mathbf{y1} - \mathbf{y2})^2}$$

Table 2 Distance matrix

	A	B	C	D	E
A	0	4	11.7	20	21.4
B	4	0	8.1	16	17.8
C	11.7	8.1	0	9.8	9.8
D	20	16	9.8	0	8
E	21.4	17.8	9.8	8	0

Table 3 Min Distance matrix

	A	B	C	D	E
A	0	4	8.1	9.8	9.8
B	4	0	8.1	9.8	9.8
C	8.1	8.1	0	9.8	9.8
D	9.8	9.8	9.8	0	8.0
E	9.8	9.8	9.8	8.0	0

Table 4 Max Distance matrix

	A	B	C	D	E
A	0	4	21.4	21.4	21.4
B	4	0	21.4	21.4	21.4
C	21.4	21.4	0	8.10	8.10
D	21.4	21.4	8.10	0	8.0
E	21.4	21.4	8.10	8.0	0

Using original distance matrix as X coordinates
And min and max distance matrix as Y
Coordinates and find relation between then

$$\mathbf{r(X,Y)} = \frac{\mathbf{N\sum XY - (\sum X)(\sum Y)}}{\sqrt{\mathbf{(N\sum X^2 - (\sum X)^2)\,(N\sum Y^2 - (\sum Y)^2\,)}}}$$

Table 5 Comparison between single linkage and complete linkage

Methods	Relational value
Single Linkage (MIN)	0.6159
Complete Linkage (MAX)	0.7196

7. ARCHITECTURE OF PROPOSED PPROACH

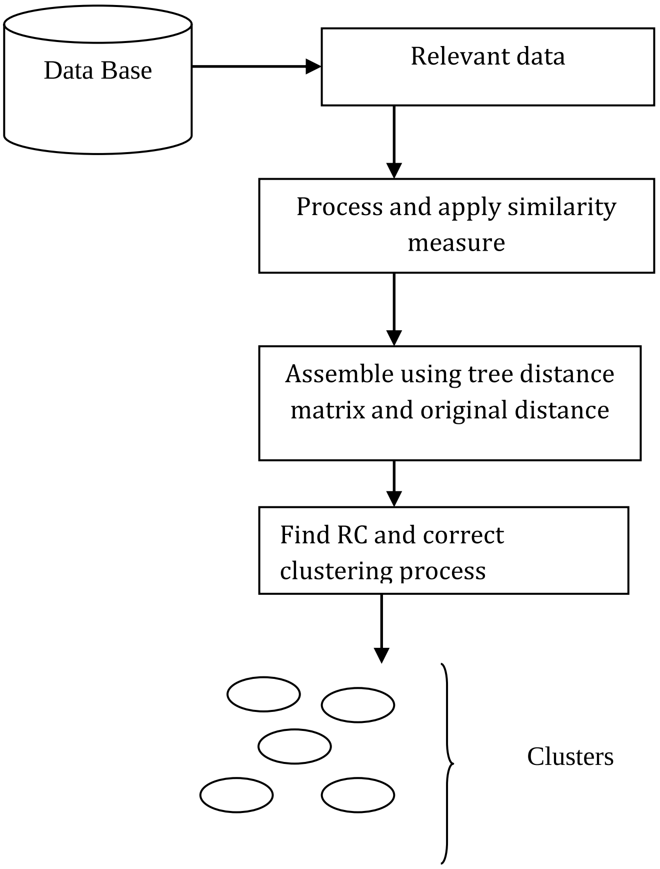
Data Base
Relevant data
Process and apply similarity measure
Assemble using tree distance matrix and original distance
Find RC and correct clustering process
Clusters

Figure 3 Architecture of proposed approach

8. EXPERIMENTAL ANALYSIS

We calculate the performance of recommended algorithm and compare it with single linkage, complete linkage and average linkage methods. The experiments were performed on Intel Core i5-4200U processor 2GB main memory and RAM: 4GB In built HDD: 500GB OS: Windows 8. The procedures are applied in using C# Dot Framework Net language version 4.0.1. Synthetic datasets are used to evaluate the performance of the algorithms.

We have taken 50 objects in two dimensional plans. Maximum value for X coordinated, 100 and Maximum value for Y coordinated is also 100. User can give the coordinated value for any object between 0 to 100 for pair of X and Y. SQL Server R2 (2008) to store our database. Database contain three attribute first is name or number of the object, second X coordinated value and third is Y coordinated value.

This snapshot displays the merging process for clustering. When user click on calculate button the accuracy value is shown in the text box. From the click for merge button user can see the step by step merging of clusters.

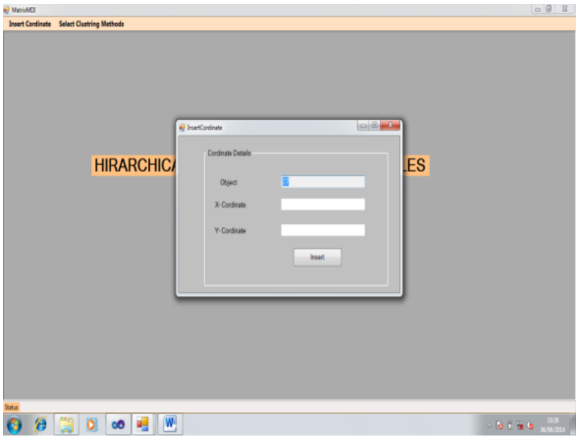

Figure 4 objects input

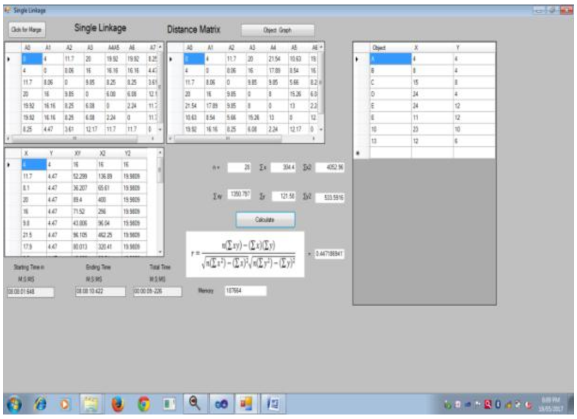

Figure 5 working of complete linkage

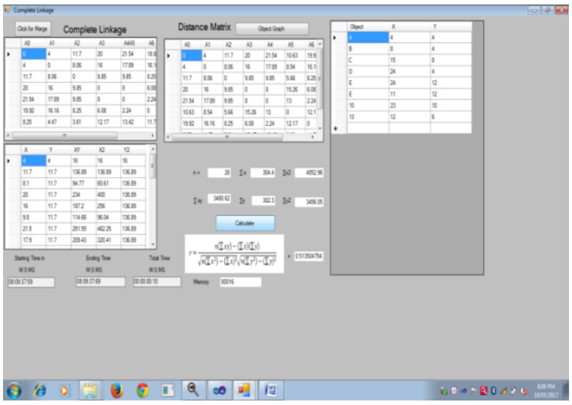

Figure 6 working of complete linkage

9. GRAPH AND ANALYSIS

Table 1 show number of objects and accuracy for single linkage and complete linkage

Table 1 accuracy with different objects

Number of Objects	Single Linkage	Complete Linkage
50	0.395674	0.63396
100	0.355668	0.538981
150	0.282241	0.424759

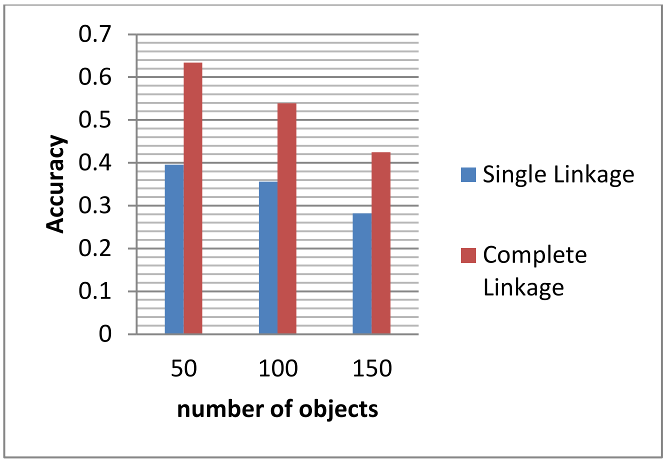

Figure 7 Comparison with Number of objects and accuracy

10. CONCLUSION AND FUTURE WORKS

There are numerous procedures and approaches have been developed for clustering problem. But problem are always arises for finding a new algorithm and process for extracting knowledge for improving accuracy and efficiency The most popular agglomerative clustering procedures are Single linkage ,Complete linkage , Average linkage and Centroid .

All of these linkage algorithms can produce totally dissimilar results when used on the same dataset, as each has its specific properties. The complete-link clustering methods usually produce more compact clusters and more useful hierarchies than the single-link clustering methods, yet the single-link methods are more versatile. Final conclusion is that the all methods are fine but to select a method for a given Situations it depends the nature of the objects.

In future enhancement we can also apply various other techniques for assembling clusters like neural network, fuzzy logic, genetic algorithms etc. to enhance the clustering.

REFERENCES

[1]J. Han, M. Kamber "Data mining, Concepts and techniques" Academic Press, 2003 page no 408 to 418.

[2]Arun K. Pujari, "Data mining Techniques" University Press (India) Private Limited, 2006, page no 122 to 124.

[3]D. Hand et al "Principles of Data Mining" Prentice Hall of India, 2004, pages no 185 to 189.

[4]Nachiketa Sahoo et al "Incremental Hierarchical Clustering of Text Documents May 5, 2006.

[5] Revati Raman et al. "Fuzzy Clustering Technique for Numerical and Categorical dataset" International Journal on Computer Science and Engineering (IJCSE) NCICT 2010 Special Issue.

[6] Hussain Abu-Dalbouh et al "Bidirectional Agglomerative Hierarchical Clustering using AVL Tree Algorithm". IJCSI International Journal of Computer Science Issues, Vol. 8, Issue 5, No 1, September 2011 ISSN (Online): 1694-0814.

[7] Dan Wei et al. "A novel hierarchical clustering algorithm for gene Sequences" 2012 Wei et al.; licensee BioMed Central Ltd.

[8] Elio Masciari et al. "A New, Fast and Accurate Algorithm for Hierarchical Clustering on Euclidean Distances" J. Pei et al. (Eds.): PAKDD 2013, Part II, LNAI 7819, pp. 111–122, 2013. Springer-Verlag Berlin Heidelberg 2013.

[9] J Anuradha et al "Hierarchical Clustering Algorithm based on Attribute Dependency for Attention Deficit Hyperactive Disorder" I.J. Intelligent Systems and Applications, 2014, 06, 37-45 Published Online May 2014 in MECS.

[10] Z. Abdullah et al "Hierarchical Clustering Algorithms in Data Mining" World Academy of Science, Engineering and Technology International Journal of Computer, Electrical, Automation, Control and Information Engineering Vol:9, No:10, 2015.

[11] Piyush Rai et al. "Data Clustering: K-means and Hierarchical Clustering" CS5350/6350: Machine Learning October 4, 2011

[12] Neepa Shah et al "Document Clustering: A Detailed Review". International Journal of Applied Information Systems (IJAIS) – ISSN: 2249-0868 Foundation of Computer Science FCS, New York, USA Volume 4– No.5, October 2012.

[13] Amit Kumar Kar et al. "A Comparative Study & Performance Evaluation of Different Clustering Techniques in Data Mining". ACEIT Conference Proceeding 2016.

[14] Shubhangi Pandit et al "An Improved Hierarchical Clustering Using Fuzzy C-Means Clustering Technique for Document Content Analysis" Volume 7, Issue 4, April 2017 ISSN: 2277 128X International Journal of Advanced Research in Computer Science and Software Engineering Research.

[15] Akshay Krishnamurthy et al "Efficient Active Algorithms for Hierarchical Clustering" Appearing in Proceedings of the 28 the International Conference on Machine Learning, Bellevue, WA, USA, 2011.

BIOGRAPHY

Mr.Anup D.Sonawane.
PG Scholar,SVCE,Indore
M.P. India