Volume 3, Issue 4, April 2013 ISSN: 2277 128X

International Journal of Advanced Research in Computer Science and Software Engineering

Research Paper

Available online at: www.ijarcsse.com

# A New Approach of Document Clustering

**Chandan Jadon**
*Department of Computer Science*
*M.Tech Student of Poornima College of Engineering, Jaipur* ,India.

**Ajay Khunteta**
*Department of Computer Science*
*(Prof.)Poornima College of Engineering, Jaipur,* India.

***Abstract— Document clustering help in organising documents in groups according to their similarity of contents. This paper presents the study of various clustering techniques. In particular K-means [3], Agglomerative Hierarchical Clustering. In addition to the various clustering techniques this paper also discusses about various document-representing techniques in graph .In particular Vector Space Model [2, 18] and Matrix Representation [2]. After studying all these things, we created a new approach of clustering algorithm and also a new representation technique of documents. We compared our results with K-means Algorithm and found that our approach is giving good results.***

***Keywords - Document Clustering, K-means, Vector Space Model, Agglomerative Hierarchical Clustering.***

## I. INTRODUCTION

Unbounded growth of electronic data has created a need to develop advanced technologies in data mining field. Document clustering involves dividing a set of documents into clusters, sharing common properties and keywords. The documents within each group should exhibit a large degree of similarity and the similarity among different clusters should be reduced[5,8].Basically there are two types of clustering techniques- "Hierarchical" and "Partitioning". Mostly we classified document-clustering algorithm into these two groups. Hierarchical Clustering often uses a tree like structure called a dendrogram which displays cluster and sub clusters relationships and the order in which the clusters were merged (Agglomerative) or split (PDDP)[3,4,6,16,17].The Partition Clustering method partition a group of documents into a set of non-overlapping clusters. Hierarchical Clustering method often define as a better quality clustering approach, but it is limited because of its quadratic time complexity and also it does not contain any provision for the reallocation of entities [1,12]. In research, it has been found out that the Partitioning Clustering Method is suited for clustering a large document dataset due to their linear time complexity. In addition to the document clustering algorithm the output of the clustering algorithm also depends on the document representation technique .So far we have two type of document clustering technique one is "Vector Space Model" and "Matrix Representation" [2].Vector Space Model is used to represent documents and web pages. It represent the collection of document in form of metrics containing row as number of document and column number as the number of terms representing the overall documents, containing the weight of the term and according to this similarity measures has been taken out. The Matrix Representation Technique represents each document as matrix Mi and the number of rows showing segment within documents. Both Vector Space Model and Matrix Representation Technique Require to check the similarity measures between each and every term present in collection of document. Here in this paper we are suggesting new clustering approach as well as document representing approach and will compare them with k-means clustering algorithm. Section II will provide an outline of document pre-processing. Section III discusses about Various Clustering Algorithm. Section IV discusses about various document representation technique. Section V discusses the measures for cluster quality that will be used as the basis for comparing different type of clustering. Section VI gives new approach algorithm and flow chart .VII gives details of the test data, and the result. The final conclusion and future work is given in section VIII.

## II. Pre-Processing Of Documents

The pre-processing is a process to optimise the list of terms that act as the keywords list. The concept of pre-processing is used to prune all character and term from the document with poor information. Removing stop words starts the first process. The stop words are words that carry no information and meaning less when we use them as search term (keyword) (i.e., Pronouns, prepositions, Conjunctions, Punctuations). Stop words may be eliminated using a list of stop words .In our experiment we are using the list of stop words. These stop words is replace by space in document we get a document with very few terms that have some meaning in search.The second process is stemming a word. Usually we have variants of words for example helps, helping, helped are the variants of word help having similar meaning. So in the process of stemming we actually find main word and replace its variant by the main word to know its occurrences in the document, removing their affixes does the stemming[5,15,17,18].

## III. Clustering Algorithm

*A. K-means clustering algorithm*

In K-means clustering algorithm we select K initial centroids [3, 4, 7], where K is a user's parameter, according to clusters desired. Each point is then assigned to the closest centroid, and each group of points assigned to centroid is a

cluster .The point is assigned to the centroid according to the distance between centroid and the point, which is calculated by some distance measure formula e.g. Euclidean distance. The mean of distances calculated to find again possible centroid .We repeat the assignments and updates until the centroids remain the same [1, 7,14, 15, 17].

**K-means Algorithm:-**

1. Select K points as temporary centroids
2. **Repeat**
3. Form K cluster by assigning points to its closest centroid (having smallest distance)
4. Compute the centroid vector $C_j$ of each cluster.(By taking median of all the points found)

$$C_j = \frac{1}{n_j} \Sigma d_{ij} \quad \text{hère } i=1 \text{ to } n_j \qquad (1)$$

Where $d_{ij}$ denote the document vector that belong to cluster $G_j$ ,$C_j$ stands for centroid vector; $n_j$ is the number of document vectors that belong to cluster $G_j$.

5. **until** centroids do not change.

*B .Agglomerative Hierarchical Clustering Algorithm*

1. Initially each point $x_1 . . . x_n$ is in its own cluster $C_1, . . . , C_n$.
2. Evaluate distances between clusters
3. Merge the nearest clusters; say $C_i$ and $C_j$ into one.

Repeat steps, until there is only one cluster left:

The result is a cluster tree. We can divide the tree at any stage to produce different clusters.

But sometimes we need to set a threshold value. The threshold value must be selected in such a way, if the distance between two clusters is more than threshold value, they do not come in same cluster. Finally we will get one cluster [4, 18].

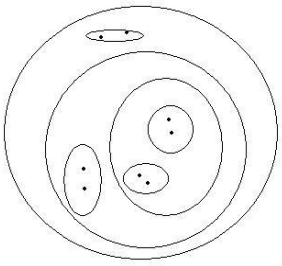


**Fig. 1**

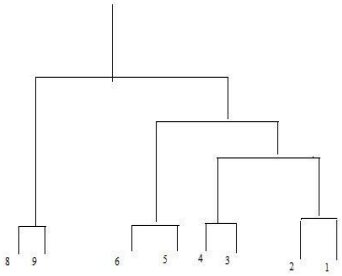



**Fig. 2**

## IV. Document Representation Technique

*A.Vector Space Model*

Vector Space Model is used to represent documents. After preprocessing of text document, it maps the document to a multi-dimensional vector with one dimension per "term" [13]. Such vectors tend to be very big. They consume time in calculating with unnecessary term also. In this model each document $D_i$, is considered to be a vector. $D_i$, is represented by m terms, $D_i = (t1, t2, t3 \ldots tm)$, where $t_j$ is the frequency of $j^{th}$ term in the document $D_i$. A collection of n documents is described by m terms can be represented by n x m matrix A, referred to as document term matrix. It is a matrix whose rows correspond to documents and whose columns correspond to the weighted terms in documents [11].Now the similarity measures in Vector Space Model presentation in graph can be calculated as follows. The cluster methods do not work directly with text documents. Instead, they calculate similarity between all pair of documents needed to build

clusters. In Vector Space Model similarity between various document can be calculated as follows, for example consider two documents $D_a$ and $D_b$ that have two keywords along x and y dimensions.

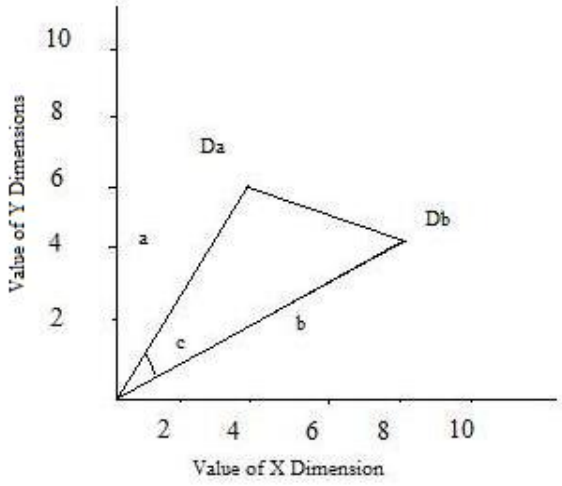



**Fig. 3**

The distance from the origin indicates the strength of the dimension in the document. The similarity between $D_a$ and $D_b$ can be calculated by using Euclidean distance, Jacquards coefficients or cosine measures etc. For more keywords there must be more dimensions [13].

*B.Matrix Representation of Documents*

In matrix representation method each document represented as a number of segments and each segment have limited number of keywords which can be mapped to the clusters. In matrix representation model rows represent different terms and columns represent cohesive segments. Matrix representation model actually represent document as a matrix instead of a vector. The basic concept is that each document is divided into more than one segment. Segments are represented by terms involved in it.It is good and we go to the deep level of document but the calculation become too much high [2].

## V. Evaluation Of Cluster Quality

We use two methods to measure the cluster quality and "goodness". One is internal quality measure clustering formalize the goal of attaining high intra-cluster similarity (documents within a cluster are similar) and low inter-cluster similarity (documents from different clusters are dissimilar) and other is external quality measure [1, 3, 4, 7, 12, 17].One of the ways of measuring the quality of a clustering solution is cluster purity. Other external measure is entropy, which provides a measure of "goodness" for un-nested clusters or for the clusters at one level of a hierarchical clustering. Another measure is the F-measure, which is more useful in measuring the quality of a clustering.

*A.Purity*

Purity is a simple and transparent evaluation measure. Let there be k clusters (the k in k-means) of the dataset n and size of cluster $C_i$ be $|C_i|$. Let $|C_i|_{class=j}$ is number of items of class j given to cluster i. Purity of cluster is given by

$$\text{Purity } (C_i) = \frac{1}{|C_i|} \max_j \left( |C_i|_{class=j} \right) \quad (2)$$

The overall purity of a clustering result can be expressed as a sum of all cluster purities.

$$\text{Purity} = \sum_{j=1}^{k} \frac{|C_i|}{|n|} \text{ purity } (C_i) \quad (3)$$

In general, larger the value of purity betters the solution.

*B. Entropy*

The entropy of a given cluster $C_i$ is defined by:

$$E_i = -\sum_j P_{ij} \log P_{ij} \quad (4)$$

Where Pij is the purity of class j in cluster i. A cluster has zero entropy if the all the documents in the cluster have the same label, otherwise it has a positive entropy.

The total entropy is the weighted average of the individual cluster entropies and is given by:

$$E = \sum_i \frac{n_i}{n} E_i \quad (5)$$

Lower the entropy of a cluster, better the quality.

*C. F-Measure*

The other external quality measure is the F measure, it combines the precision and recall values from information. We then calculate the recall and precision of cluster for every class.
Precision (p) and recall (r) compares each cluster i with each class j in the classification:

$$P_{ij} = \frac{n_{ij}}{n_i} \quad (6)$$

$$r_{ij} = \frac{n_{ij}}{n_j} \quad (7)$$

Where $n_{ij}$ is the numbers of documents from class j in cluster i, $n_i$ is the number of document in cluster i and $n_j$ is the number of documents in class j. The corresponding value under the F measure is

$$F_{ij} = \frac{2 * r_{ij} * p_{ij}}{r_{ij} + p_{ij}} \quad (8)$$

This approach is to take the best cluster i (the one with the highest F score Fj ) for each topic j as the cluster matching that topic and perform a weighted average of these best F values:

$$F_j = \max_i \{ F_{ij} \} \quad (9)$$

$$F = \sum_j \left( \frac{nj}{n} \right) * F_j \quad (10)$$

Where n is the number of texts in the whole text set.

## VI. New Approach

*A. Representation*

In new approach of document representation we have a document - term matrix .The rows represent the documents and the columns represent the terms number(which are fix for each term).The terms are arranged in such a manner ,the term number is first in the list whose weight is highest and they are arranged in decreasing order of weight(frequency in document)[9,10].For example let we have 10 documents and we have to cluster them according to the given 20 terms.The twenty terms that has been selected have given fix numbers e.g. term1,term 2, term 3-----term 20.Now if in document $D_1$, term 2, term 5 and term 12 has the highest weight. They will be arranged in decreasing order then $D_1$'s vector is represented as $D_1$[ 2, 14, 12].

We prefer three dimensions as we assume that three keywords are sufficient in finding the cluster of the document to which it belongs.The n document and 20 key words matrix is obtained as n x 3 matrix or we can increase the columns to increase the accuracy.Our Approach of algorithm is saying that whatever number of cluster we are going to make, we need to divide the number of keywords in different type of classes(domain).As we have given each keyword a number, we represent each document Di as vector of three numbers .$D_i=\{p_i, q_i, r_i\}$ , where $p_i$, $q_i$, $r_i$ are the numbers between 1 to m, m is number of keywords, $p_i$, $q_i$, $r_i$ representing the keywords number in such a manner that weight($p_i$)>weight($q_i$)>weight($r_i$) and $p \neq q \neq r$.Now the actual algorithm begins with grouping the documents in different cluster.

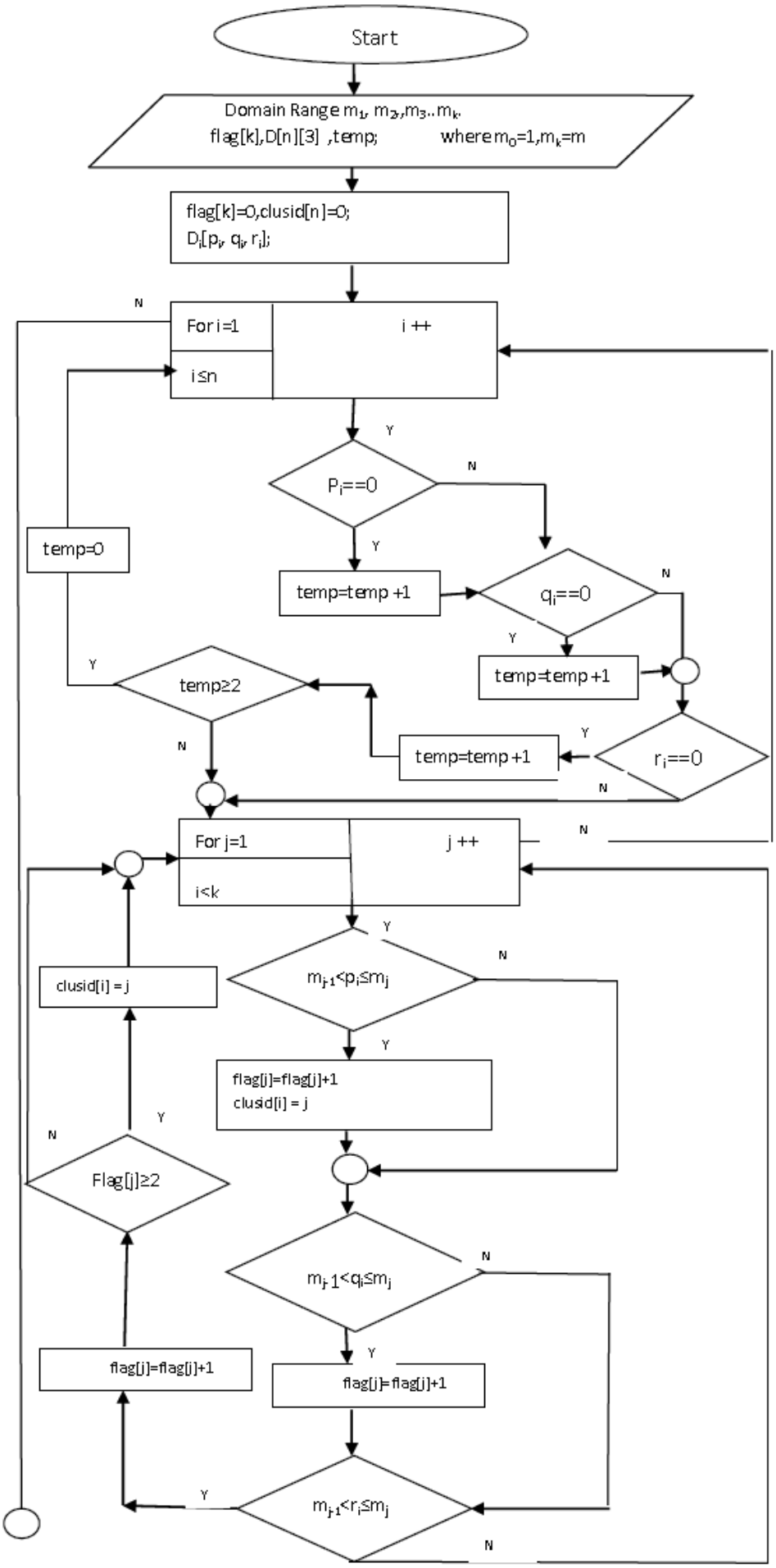
B. *Algorithm*

Steps :

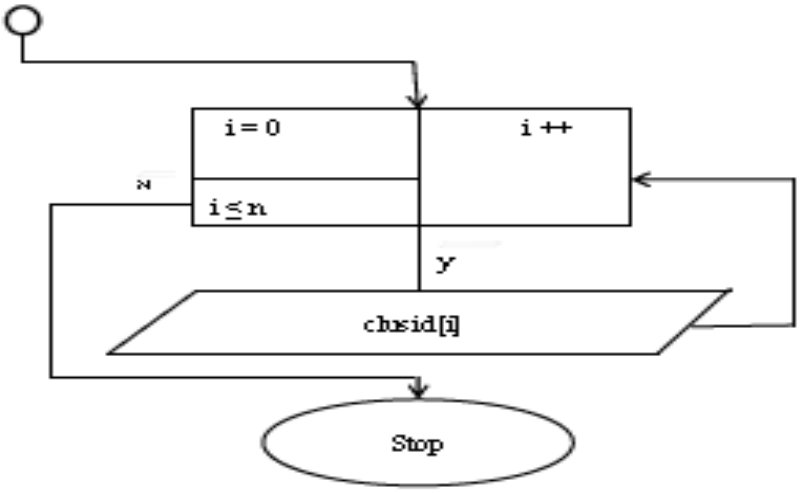
1. Divide the keywords in specific k domain e.g. [1 to $m_1$,$m_1$+1 to $m_2$, $m_2$+1 to $m_3$, … $m_{k-1}$+1 to m. Where $m_1$,$m_2$,$m_3$ are integers $m_1<m_2<m_3 \ldots m_{k-1}<m$.
2. Initialise flag 1,flag 2…flag k =0.
3. 3. for $D_i$ Document , i =1 to n, n is number of document.
   - 3.1 $D_i=\{p_i, q_i, r_i\}$ ; $p_i, q_i, r_i \in \{1 \text{ to } m\}$
   - 3.2 If $p_i$ && $q_i$ or $q_i$ && $r_i$ or $p_i$ && $r_i$ or $p_i$ && $q_i$ && $r_i$ ,belongs to the same domain where j(1<=j=<k)
     Then flag j=2 or 3
     And
     $D_i$ set to the cluster j
   - 3.3 Else if $p_i$, $q_i$, $r_i$ all belongs to different domain
     $D_i$ set to the cluster j where j is the domain to which $p_i$ belongs
   - 3.4 If in $p_i$, $q_i$, $r_i$ any two or all three not belongs to any domain then $D_i$ is an outlier.
4. Repeat step 3 till all documents get clustered.

*C.Flow Chart*






## VII. Result

A. *Test Dataset*

We have taken 140 documents and 4 classes as Computer Science, Electronics and electrical Engineering, Mechanical Engineering, Civil Engineering. We collected the keywords for each specific branch. E.g. Table I. Is showing only ten keywords for each specific branch. Originally we have taken 67 keywords in Computer Science,64 keywords in civil, 49 in electrical and electronics,56 in mechanical. As table showing 1 to 10 keyword numbers for computer science, 11 to 20 numbers for civil, 21 to 30 for Electrical and Electronics and 31 to 40 for mechanical words, so now these are four domains. If we apply our algorithm by getting document term matrix, all three terms must be in descending order of occurrences, we see through result it represent document in best way to particular class.

**TABLE I**

| **key no.** | **Computer sc.class** | **key no** | **Civil class** | **keyno** | **Electrical & Electrical class** | **keyno** | **Mechanical** |
|---|---|---|---|---|---|---|---|
| 1 | Algorithm | 11 | aggregate | 21 | Amplifier | 31 | Automobile |
| 2 | Array | 12 | arch | 22 | Antenna | 32 | Axle |
| 3 | C++ | 13 | bridge | 23 | attenuation | 33 | Chain |
| 4 | Cache | 14 | building | 24 | circuit | 34 | Clutch |
| 5 | Computer | 15 | cement | 25 | electronic | 35 | Engine |
| 6 | Database | 16 | clay | 26 | electrostatic | 36 | Fluid |
| 7 | Structures | 17 | dam | 27 | microprocessor | 37 | Lubrication |
| 8 | file | 18 | dome | 28 | microwave | 38 | Machine |
| 9 | Linked list | 19 | road | 29 | Radar | 39 | Pattern |
| 10 | Linux | 20 | rocks | 30 | semiconductor | 40 | Robot |

A. *Result*

The Fig.(4) is showing four clusters over 236 keywords calculated according to our algorithmic approach.

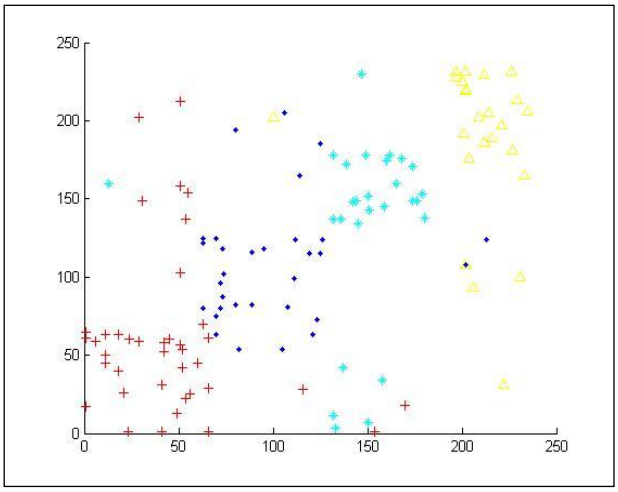


**Fig. 4**

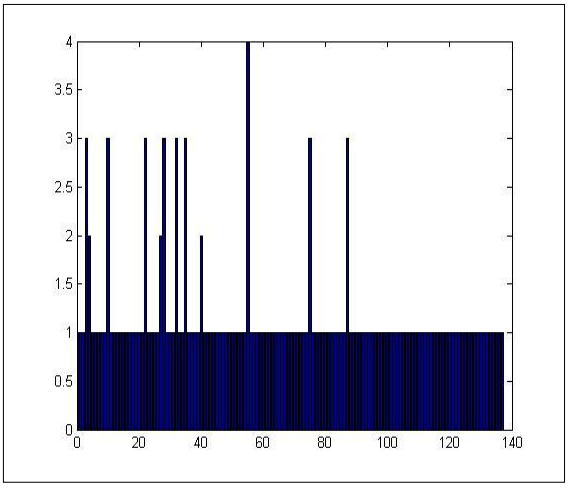



**Fig. 5 Clustering by k-means algorithm**

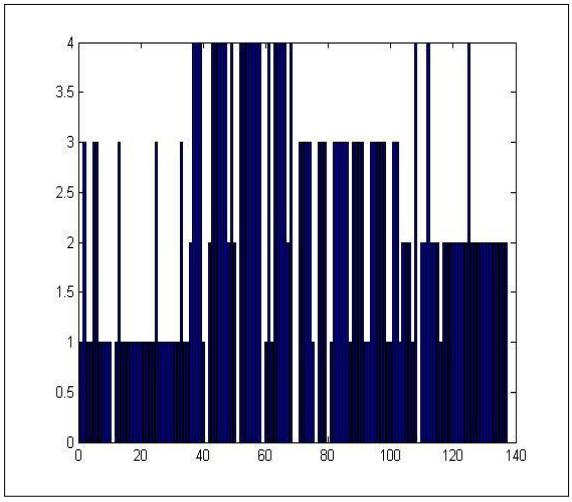



**Fig. 6 Clustering by New Approach**

Fig.(5) is showing clustering by k means on our keywords and our data we are not getting a good result as it has taken maximum number of document in one cluster.
Fig.(6) is showing clusters and the clusters are good, getting all the documents in one clusters belonging to the same class. As we have compared our result with k means algorithm applied on same set of document and key word . Table II. is showing purity, entropy ,F measure for k means as well as for our algorithm.

**TABLE II**

| S. No. | Algorithm | Documents | Keywords | Class | Purity | Entropy | F measure |
|---|---|---|---|---|---|---|---|
| 1. | K-means | 140 | 260 | 4 | .32846 | .18369 | .35769 |
| 2. | New approach | 140 | 260 | 4 | .83211 | .040839 | .80555 |

## VIII. Conclusion And Future Work

After doing experimental studies on both the algorithm we get a conclusion that according to the need we should create new algorithms over previous algorithm and focus on getting less time complexity if possible. If domain and keywords are known, this approach is good and it is very helpful in statistical analysis of employees, students etc. In future work we are planning to make our algorithm more general by performing more improvement over it.

**Refrences**

[1] V. Mary Amala Bai and Dr. D. Manimegalai, "An Analysis of Document Clustering Algorithms", 2010 IEEE
[2] Xufei Wang, Jiliang Tang and Huan Liu, "Document Clustering via Matrix Multiplication" 2011 11th IEEE InternationalConference On Data Mining.
[3] Shi Zhong,"A k-means algorithm to improve the Efficiency Using Normal Distribution Data Points",(IJCSE) International Journal on Computer Science and Engineering, 2010.
[4] K. Rajendra Prasad and Dr. P.GovindaRajulu "A Survey on Clustering Technique for Datasets using Efficient Graph Structures" International Journal of Engineering Science and Technology Vol. 2 (7), 2010, 2707-2714
[5] Anil K. Jain, "Pattern Recognition Letters", Journal Elsevier, Pattern Recognition Letters 31 (2010) 651– 666.
[6] Xufei Wang and Liang Tang ," Cluster Analysis, Basic Concepts and Algorithms", 2011 11th IEEE International Conference on Data Mining.
[7] D.Napoleon and P.Ganga Lakshmi,"An Enhanced k-means algorithm to improve the Efficiency Using Normal Distribution Data Points". (IJCSE) International Journal on Computer Science and Engineering Vol. 02, No. 07, 2010, 2409-2413
[8] Atika Mustafa, Ali Akbar, and Ahmer Sultan"Knowledge Discovery using Text Mining: A Programmable Implementation on Information Extraction and Categorization", International Journal of Multimedia and Ubiquitous Engineering Vol. 4, No. 2, April, 2009.
[9] I.C. Mogotsi, News analysis through text mining: a case study : VINE: The journal of information and knowledge management systems Vol. 37 No. 4, 2010 pp. 516-531.
[10] Chen-Huei Chou, Atish P. Sinha, Huimin Zhao"A Hybrid Attribute Selection Approach for Text Classification, Journal of association for information system", Volume 11, Issue 9, pp. 491-518, September 2010.
[11] Shobha S. Raskar, D. M. Thakore "Text Mining and Clustering Analysis", IJCSNS International Journal of Computer Science and Network Security, VOL.11 No.6, June 2011.
[12] Michael Steinbach, George Karypis and Vipin Kumar," A Comparison of Document Clustering Techniques" Department of Computer Science and Egineering, University of Minnesota Technical Report #00-034.

[13] Mrs .S.C.Punitha and Dr.M.Punithavalli, "A Comparative Study to Find A Suitable Method for Text Document Clustering", International Journal of Computer Science & Technology(IJCSIT)Vol 3,No 6 December 2011.

[14] Kardi Teknomo,PhD ,"K-Means Clustering Tutorial" Teknomo ,Kardi,K-means Clustering Tutorials .http://people.revoledu.com\kardi\tutorial\kmeans\.

[15] Shi Zhong," Efficient Online Spherical K-means Clustering".

[16] Berkhin, P., 2002. "Survey of clustering data mining techniques.Accrue Software Research Paper."

[17] Manu Konchandy,"Text Mining Application Programming", Publisher: Cengage Learning. ISBN13: 9788131502471

[18] Book: Information Retrieval, Algorithms and heuristics by David A.Grossman and Ophir Frieder.Published by Springer International.