ISSN:0975-9646

Implementation of Text clustering using Genetic Algorithm

Dhanya P.M#, Jathavedan M*, Sreekumar A*

#*Department of Computer Science*
Rajagiri School of Engineering and Technology, Kochi, India, 682039
**Department of Computer Applications*
Cochin University of Science and Technology, Kochi, India, 682022

***Abstract*— Text clustering is an important area of interest in the field of Text summarization, sentiment analysis etc. There have been a lot of algorithms experimented during the past years, which have a wide range of performances. One of the most popular method used is k-means, where an initial assumption is made about k, which is the number of clusters to be generated. Now a new method is introduced where the number of clusters is found using a modified spectral bisection and then the output is given to a genetic algorithm where the final solution is obtained.**

***Keywords*— Cluster, Spectral Bisection, Genetic Algorithm, k-means.**

I. Introduction

A lot of clustering methods have been evolved during the past years. In Manjot Kaur et. al.[1], k-means algorithm is used to cluster web documents. The disadvantage of k-means is that the success depends on k, the number of clusters selected. The value of k is assumed to be 3, 4 etc. Actually the number of clusters can be less than or more than k which we select. Euclidean distance between the documents is considered as the similarity criteria. In Rakesh Chandra et. al.[2], both k-means and k-medoid algorithms are used .In case of k-means the centroid is modified by taking the mean of the cluster, in k-medoid the most centrally placed object of the cluster is taken as the medoid. In Banjamin Fung et.al. [3], frequent item sets are considered for reducing the dimensionality of the document .Here cluster trees are formed after child pruning and sibling merging. In Yunsha ,Guonying Zhang et.al. [4], lexical graph is created where the edges are marked with correlation grade of the words. The class name will be the nodes with the higher degrees. In Houfeng Ma et.al [5], they use an incremental learning method based on Bayes probability. In Xi Guo et.al.[6],clustering is done by converting text in to situation vectors which actually consists of SVi =(P,A,T,S) where P consists of noun phrases, A consists of verb phrases and predicates, S means location relevant noun phrases and T is the Time relevant information in the Text. From this a cognitive situation matrix is constructed. Now the correlation between cognitive situation matrices is found out for clustering.

The method suggested by Jinzhu Hu et.al.[7] ,the majority of the noisy words are removed and an initial clustering is made and then precision of the clustering is improved by reducing the dimension of the feature vector. In the method proposed by Zhenya Zhang et.al.[8], fitness function is defined as disagreement of C1 and C2 ,d(V,C1,C2) where C1 and C2 are two clustering divisions of dataset V and in method [9] proposed by him fitness function is modeled by taking in to account the correlation matrix. In the method [10] clustering is based on frequent term sets . Each selected frequent term set is the description of a cluster. In method coined by Ranjana Agrawal et.al.[11],cosine similarity matrix is used and a comparison is made with the k-means algorithm. In Tayfun Dogdas et.al.[12], Expectation Maximization and Multidimensional projection methods are used for clustering documents. In Sarnovsky et.al.[13], growing hierarchical self organizing maps are used .

The main aim of this paper is to device a good clustering method for sentences in a document and then applying the same for identifying the subtopics in a document. Similar sentences are grouped together to form a cluster, so that they pertain to the same subtopic. The method consists of Phase 1 and Phase 2 where the Phase 1 is clustering the sentences based on modified repeated spectral bisection. Here an initial clustering is done based on the sign of elements of the eigen vectors of the similarity matrix [14].Phase 2 is clustering sentences based on the genetic algorithm [15], where chromosomes are created which are actually solutions to the clustering problem.

II. System Architecture

Here the sentences from document to be processed are tokenized and the stop words are removed. The remaining set of words are stemmed as in Fig .4 and are used to construct the adjacency matrix A. The diagonal matrix D of vertex degrees is used to find out the similarity matrix C which is equal to D-A. From the eigen values of the similarity matrix, the one which is near to $\boldsymbol{\lambda}_{\mathbf{d}} = \mathbf{m} / \mathbf{c}$ is selected, where m is the total number of edges and c is the sum of vertex degrees. The sentences are divided in to two groups based on the sign of the elements in the eigen vector corresponding to this eigen value. This is repeated until there is no change in sign. Now a second level clustering is carried out by creating a chromosome, where it has a header and data part. The header stores the Meta data like the number of clusters k, and the number of elements in each cluster n_k. The data portion of the chromosomes in the initial population is subjected to crossover, mutation and selection of unique chromosomes. The fitness value is now applied to each and every chromosome. The process is iterated 10 times and the chromosome with the highest fitness is selected as the solution to clustering. The system architecture is given in Fig . 1.

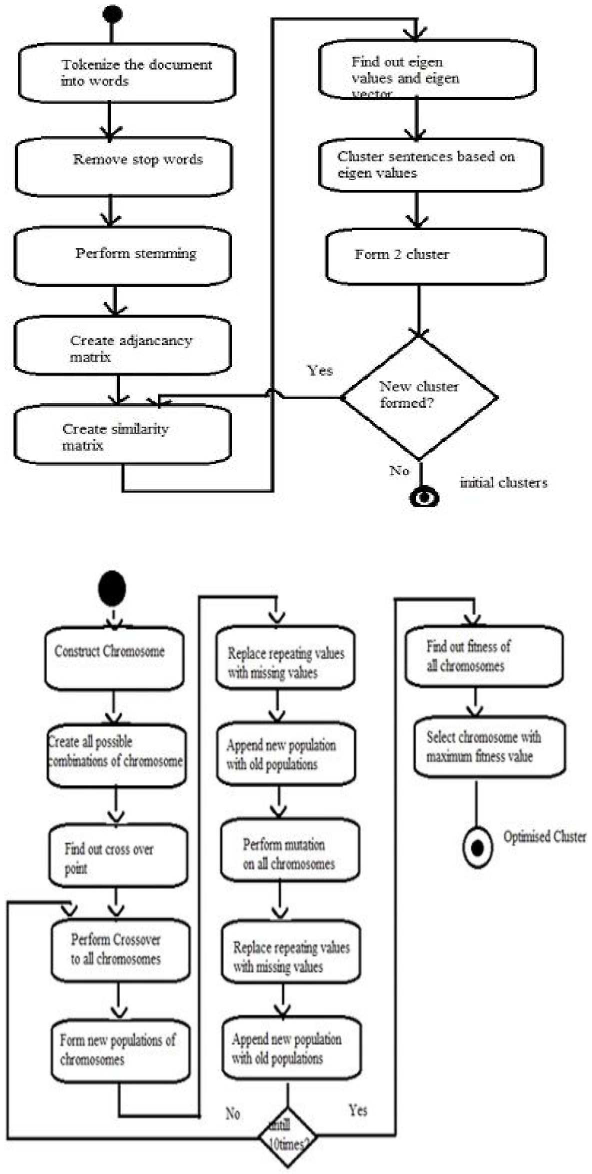

Tokenize the document into words
Remove stop words
Perform stemming
Create adjancancy matrix
Create similarity matrix
Find out eigen values and eigen vector
Cluster sentences based on eigen values
Form 2 cluster
Yes
New cluster formed?
No
initial clusters
Construct Chromosome
Create all possible combinations of chromosome
Find out cross over point
Perform Crossover to all chromosomes
Form new populations of chromosomes
Replace repeating values with missing values
Append new population with old populations
Perform mutation on all chromosomes
Replace repeating values with missing values
Append new population with old populations
No
untill 10times?
Yes
Find out fitness of all chromosomes
Select chromosome with maximum fitness value
Optimised Cluster

Fig. 1 System Architecture

A. Algorithm :Phase 1

1. text[]= sentence from document.
2. words[]= words of text[].
3. Compare words[] with stop words[].
4. Remove matching words.
5. for i = 1 to n stem (word $_i$).
6. Create A= adjacency matrix.
7. Create D= Diagonal matrix.
8. Find C= D – A.
9. Find eigen values and eigen vectors of C.
10. Select $\lambda_d = m / c$; m= total number of edges and c= total vertex degree.
11. Select eigen value near to $\boldsymbol{\lambda_d}$.
12. Select eigen vector corresponding to the eigen value.
13. Sentences corresponding to positive and negative elements in the eigen vectors are divided in to two groups.
14. Repeat steps 6 to 13 until no change in sign.
15. k= no of clusters in phase1

This method and the result analysis of this method is already discussed in [1] . But the main disadvantage is that the algorithm is not always stable. This method is strengthened by applying genetic algorithm. The number of clusters k is given to phase II where a chromosome is being created as the following

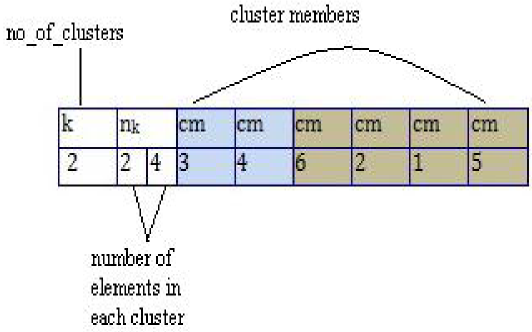


Fig. 2 Sample Chromosome with header

The chromosome shown as Fig. 2 has the header portion which stores the number of clusters and the number of elements in each cluster.

B. Genetic Algorithm : Phase 2

1. Create ST $_{n \times m,}$ the sentence term matrix where n= number of sentences and m =number of terms in the document.
2. ST [i, j] = 0, if term $_j$ does not occur in sentence $_{i}$, ST [i, j] = 1 , if term $_j$ occurs in sentence $_i$
3. Remove the header portion of the chromosome and make all the combinations of the cluster member part of the chromosome.
4. P_T has n! number of chromosomes
5. P_0 = initial population = random selection from P_T.
6. Select crossover point as n/2, where n is also the length of the chromosome.
7. Perform crossover of all chromosomes.
8. Replace repeating elements with missing elements in all chromosomes. This will create the new population Pc.
9. Add P_0 + P_c. Remove the duplicating chromosomes.
10. Perform mutation of the chromosomes by subtracting each element of the chromosome from the last element and take the absolute value.
11. Replace repeating elements with missing elements . This will create the new population Pm.
12. $P_{new} = P_0 + P_c + P_m$.
13. Take unique chromosomes from P_{new}.
14. Find fitness of each chromosome using the Jaccard formula

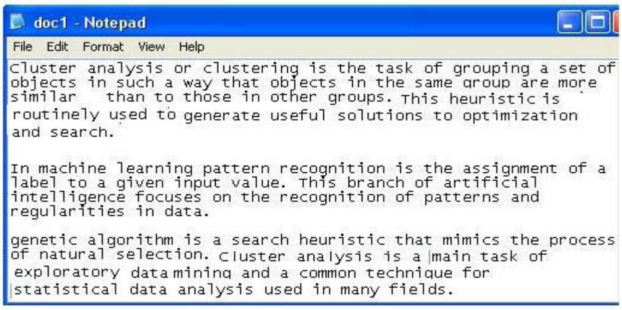
$$\text{Fitness value} = f\left(\sum_{i=1}^{k} \sum_{j=1}^{n_k-1} \sum_{p=j+1}^{n_k} \left(|A_j \cap A_p| / |A_j \cup A_p|\right)\right) \tag{1}$$

Where A_j and A_p are the rows in the Sentence term matrix.

15. Remove chromosomes with zero fitness value.
16. Do steps 6 to 15 , ten times
17. Select the chromosome with the highest fitness value which is the solution to clustering.

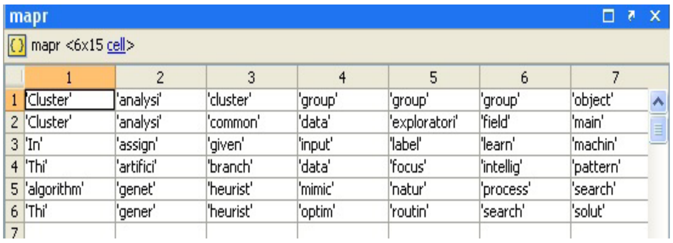
III. RESULTS

In Fig. 3, it is evident that the sentences are discussing about three topics cluster analysis, machine learning and genetic algorithm. Here the crossover probability and mutation probability is one. We are taking the crossover of all pairs of chromosomes. Decimal mutation is applied to all allele values. The resultant chromosomes would be as shown in Fig. 6. Fitness values are calculated using the formula mentioned in step 14 of Phase 2 .Fitness is calculated from the jaccard coefficient value and some samples are shown in Fig. 7. It is calculated between each pair of sentences in a cluster. These values can be obtained from the attributes in sentence term vector which is shown in Fig. 5. Higher the fitness value better is the solution.



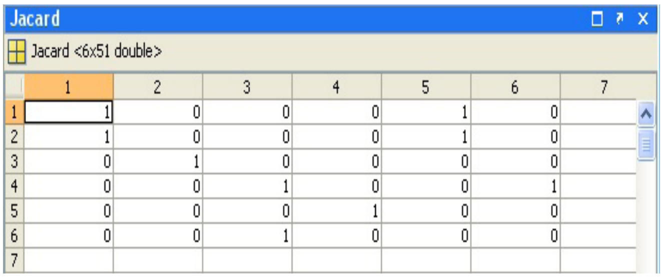
```
Cluster analysis or clustering is the task of grouping a set of
objects in such a way that objects in the same group are more
similar   than to those in other groups. This heuristic is
routinely used to generate useful solutions to optimization
and search.

In machine learning pattern recognition is the assignment of a
label to a given input value. This branch of artificial
intelligence focuses on the recognition of patterns and
regularities in data.

genetic algorithm is a search heuristic that mimics the process
of natural selection. Cluster analysis is a main task of
exploratory data mining and a common technique for
statistical data analysis used in many fields.
```

Fig. 3 Sample Document



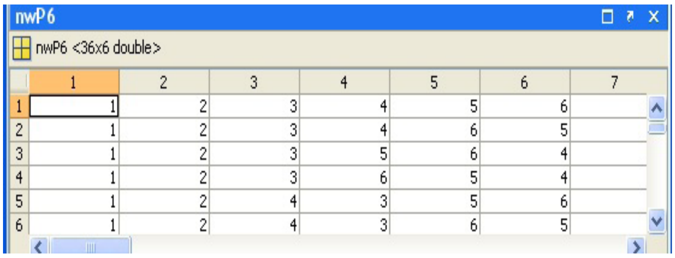
	1	2	3	4	5	6	7
1	'Cluster'	'analysi'	'cluster'	'group'	'group'	'group'	'object'
2	'Cluster'	'analysi'	'common'	'data'	'exploratori'	'field'	'main'
3	'In'	'assign'	'given'	'input'	'label'	'learn'	'machin'
4	'Thi'	'artifici'	'branch'	'data'	'focus'	'intellig'	'pattern'
5	'algorithm'	'genet'	'heurist'	'mimic'	'natur'	'process'	'search'
6	'Thi'	'gener'	'heurist'	'optim'	'routin'	'search'	'solut'
7							

Fig. 4 Stemmed words



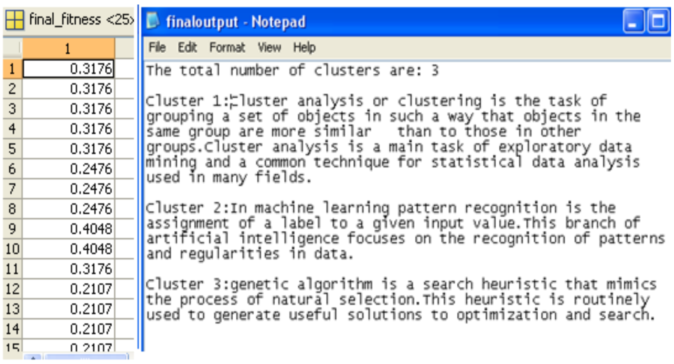
	1	2	3	4	5	6	7
1	1	0	0	0	1	0	
2	1	0	0	0	1	0	
3	0	1	0	0	0	0	
4	0	0	1	0	0	1	
5	0	0	0	1	0	0	
6	0	0	1	0	0	0	
7							

Fig. 5 Sentence Term matrix



	1	2	3	4	5	6	7
1	1	2	3	4	5	6	
2	1	2	3	4	6	5	
3	1	2	3	5	6	4	
4	1	2	3	6	5	4	
5	1	2	4	3	5	6	
6	1	2	4	3	6	5	

Fig. 6 Chromosomes after crossover, mutation.



	1
1	0.3176
2	0.3176
3	0.3176
4	0.3176
5	0.3176
6	0.2476
7	0.2476
8	0.2476
9	0.4048
10	0.4048
11	0.3176
12	0.2107
13	0.2107
14	0.2107
15	0.2107

```
The total number of clusters are: 3

Cluster 1:Cluster analysis or clustering is the task of
grouping a set of objects in such a way that objects in the
same group are more similar   than to those in other
groups.Cluster analysis is a main task of exploratory data
mining and a common technique for statistical data analysis
used in many fields.

Cluster 2:In machine learning pattern recognition is the
assignment of a label to a given input value.This branch of
artificial intelligence focuses on the recognition of patterns
and regularities in data.

Cluster 3:genetic algorithm is a search heuristic that mimics
the process of natural selection.This heuristic is routinely
used to generate useful solutions to optimization and search.
```

Fig. 7 Fitness values and Final output.

IV. RESULT ANALYSIS

The result of the above method is being compared with the k-means algorithm, which is a popular clustering method in the area of a data mining.

A) *Precision vs number of clusters.*

The graph in Fig. 8 shows the precision of both the k-means algorithm and genetic algorithm applied in clustering. Here Precision is in the scale 0 to 1. Here precision is calculated as :

$$Precision(P) = \frac{Number of Relevant Clusters}{Total number of Clusters} \tag{2}$$

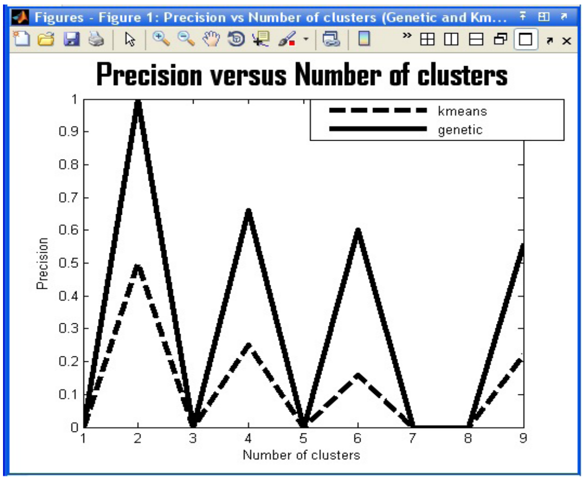


Fig .8 Precision

From the graph we can see that the precision decreases as the number of clusters increases and the genetic algorithm has a better precision than k-means.

B) Time vs number of clusters.

The graph in Fig. 9 is a plot between the time and the number of clusters calculated for both k-means and genetic algorithm. From the graph we can see that genetic algorithm takes slightly more time than k means. Well this is one of the main disadvantages of the system. Here time is calculated using the tic and toc functions of MATLAB. tic starts a stopwatch timer and when toc is executed, the variable toc will have the time elapsed since the tic function started the stop watch timer .

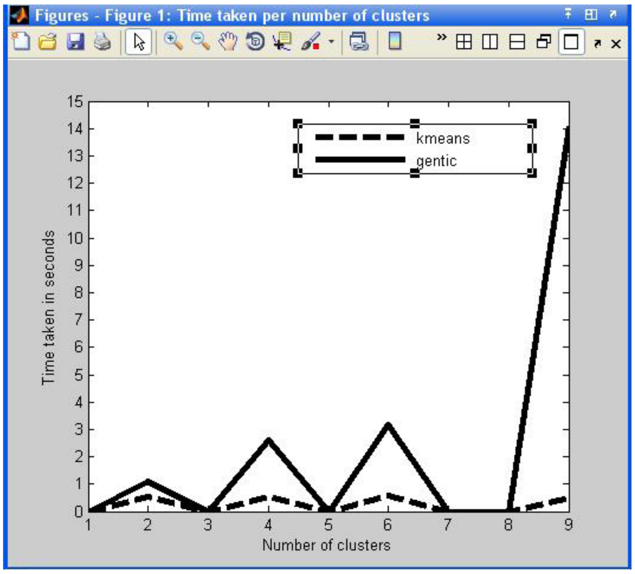


Fig .9 Time

C) Fscore vs number of clusters.

The graph in Fig. 10 is a plot between Fscore and the number of clusters. From the graph it is evident that Fscore is more for genetic algorithm when compared to k-means. Fscore is calculated as:

$$F = 2 * \frac{Precision \ * \ Recall}{Precision \ + \ Recall}$$

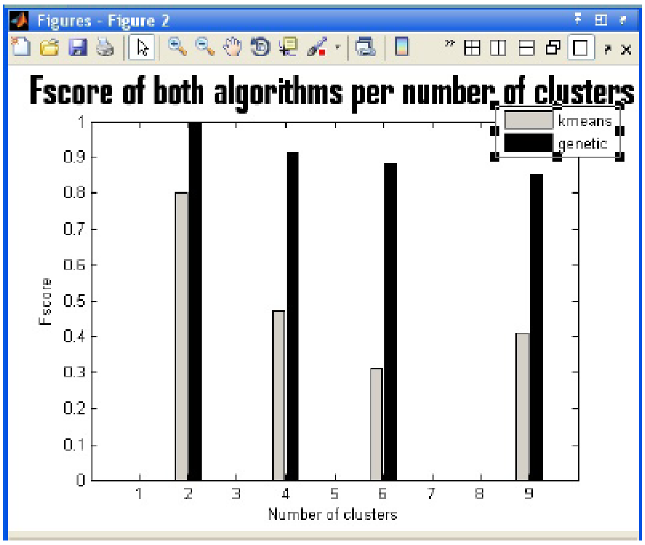


Fig .10 Fscore

D) Entropy vs number of clusters.

The Fig. 11 is a plot between the Entropy and the number of clusters. From the graph it is evident that Entropy is less for genetic algorithm, when compared to k-means. Lesser the entropy better is the algorithm. Entropy is calculated using a built function in MATLAB where the input is a sentence - term matrix. Each cell corresponds to a word in the sentence represented in the form of numbers.

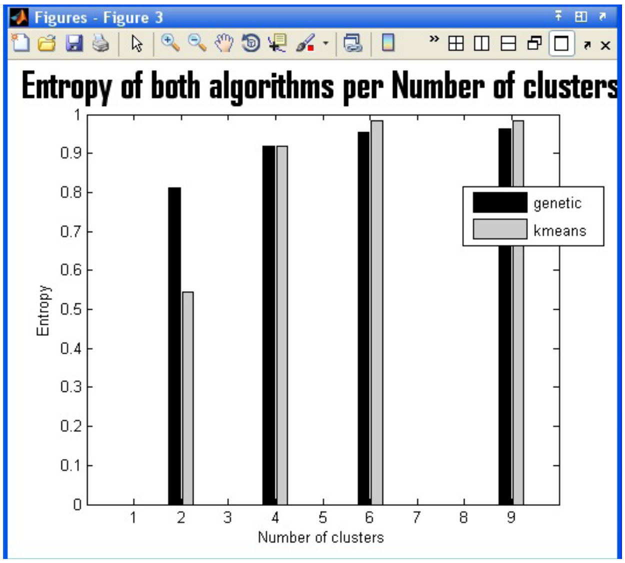


Fig . 11 Entropy

V. CONCLUSION

The paper reveals the use of a new method which successfully grouped similar sentences in a document, so that they pertain to one topic. Thus topic segmentation can be done in a single document .The use of genetic algorithm has strengthened the initial clustering .Initial clustering is a modified version of the original spectral bisection algorithm. The new method is in many ways better than popular k-means except in the case of time duration. The [3]slight increase in time of the algorithm is a disadvantage where methods have to be sorted out to reduce the time. This can be done as the future enhancement.

REFERENCES

[1] Manjot Kaur,Navjot Kaur Web document clustering approaches using k-means algorithm,International Journal of Advanced Research in Computer Science and Software Engineering,Vol 3,May 2013.

[2] Rakesh Chandra Balabantaray, Chandrali Sarma,Monica Jha,” Document clustering using k-means and k-medoids,International Journal of Knowledge based Computer Systems,Vol 1,Issue 1,2013.

[3] Banjamin C.M Fung ,KeWang ,Martin Ester, “ Hierarchical document clustering using Frequent itemsets”,In proceedings of SIAM International Conference on Data mining 2003.

[4] Yunsha , Guoying Zhang, Huina Jiang ,” Text clustering based on Lexical graph”, Fourth International Conferenec on Fuzzy System and Knowledge Discovery , 2007.

[5] Houfeng Ma, Xianghua Fan ,Jichen ,” An incremental Chinese text classification algorithm based on quick clustering, International Symposium on Information Processing , 2008.

[6] Xi Guo, Zhiqung Shao, Nam Hua, “ A hierarchical Text clustering algorithm with cognitive situation Dimensions “, Second International Workshop on Knowledge discovery and Data Mining,2009.

[7] Jinzhu Hu, Chunxiu Xiong ,Jiangbo Shu, Xing Zhou, Jun Zhu, “ A novel Text clustering method based on TGSOM and Fuzzy K-

means ",First International Workshop on Educationa Technology and Computer Science ,2009.

[8] Zhenya Zang , Hongmai Cheng " Clustering Aggregation based on genetic algorithm for Document clustering.", 2008 IEEE Congress on Evolutionary Computation (CEC 2008).

[9] Zhenya Zhang , Hongmai Cheng , " Correlation clustering based on genetic algorithm for document clustering " 2008 IEEE congress on Evolutionary Computation (CEC 2008)

[10] Qing Hi, Tingting Li, Fuzhen Zhuang," Frequent term based Peer to Peer Text clustering ",3rd International Symposium on Knowledge Acquisition and Modeling,2010.

[11] Ranjana Agrawal, Madhura Phatak," A Novel Algorithm for Automatic Document clustering ", 3rd IEEE International Advance Computing Conference IACC 2013.

[12] Tayfun DogDas , Selim Akyokus," Document clustering using GIS visualizing and EM clustering method , IEEE International Symposium on Innovation in Intelligent Systems and Applications June 2013.

[13] M. Sarnovsky ,Z Ulbuk, " Cloud based clustering text documents using GHSOM algorithm on the Grid gain platform ",8th international symposium on Applied Computational Intelligence and Informatics ,May 2013.

[14] Dhanya P.M, Jathavedan M, Application of Modified Spectral bisection for Segmenting Malayalam documents, Third International Conference on Advances in Computing and Communications, 2013.

[15] Dhanya P.M,Jathavedan M,Sreekumar A, " A proposed method for clustering Malayalam documents using Genetic Algorithm " ,Fourth National Conference on Indian Language Computing (NCILC-2014), Feb 1-2, 2014 ,Kerala ,India .