ISSN 2394-3777 (Print)
ISSN 2394-3785 (Online)
Available online at www.ijartet.com
International Journal of Advanced Research Trends in Engineering and Technology (IJARTET)
Vol. 3, Special Issue 20, April 2016

Genetic Algorithm Based Document Clustering

R. Janani

MPhil Research Scholar
Department of Computer Science and Engineering
Alagappa University-Karaikudi, India.
jananianandan38@gmail.com

***Abstract*— Document clustering is a significant domain of interest in the field of document summarization. Document Clustering is the application of cluster analysis to textual documents. Many different algorithms have been proposed to cluster documents during the past years. K-means clustering is one of the methods used for clustering documents. These methods suffer from issues and challenges like accuracy and time complexity. To overcome these limitations a novel genetic algorithm based document clustering method have been proposed. The Objective of this research work is to cluster documents for retrieved the information. This method has been implemented and experimentally evaluated using various measures. It provides improved accuracy and time complexity when experimented using real world dataset.**

***Keywords*— Genetic algorithm, Clustering, K-means, Crossover, Selection.**

I. INTRODUCTION

Information retrieval (IR) is the process of representing, storing, organizing, and allowing access to information repository. It finds and retrieves relevant text documents. Clustering is used to group similar documents and categorize documents belonging to a particular class. For efficient and effective information retrieval documents may be clustered on the basis of similar keywords.

A clustering system can be useful in web search for grouping search results into closely related sets of documents. It can improve similarity search by focusing on sets of relevant documents and hierarchical clustering methods can be used to automatically create topic directories, or organize large collections of web documents for efficient retrieval. Document Clustering is a more specific technique for unconfirmed document organization. It is generally considered to be a centralized process. Clustering methods can be used to automatically group the retrieved documents into a list of meaningful categories. It has applications in automatic document organization, topic extraction and fast information retrieval. Clustering of documents is useful for automatic organizing documents in a set or meaningful categories, and fast and efficient information retrieval. It organizes data into groups of objects called clusters. Clusters contain data instances that are similar to each other. The data instances in different clusters are very different from each other. Clustering is an unsupervised learning method. The class labels are previously unknown. It is different from supervised classification

II. RELATED WORKS

Many traditional clustering techniques has been examined during the historical times. k- means clustering is used to cluster web content[1]. The drawback of k-means clustering is to be dependent on k number of clusters had been selected. In case of k-means the centroid is altered by attractive the cluster mean, in k-medoid method cluster has most of centrally placed object as the medoid. Banjamin Fung et.al. [2], discussed to considered for reducing the dimensionally of the text using frequent item sets. First doing sibling merging and child pruning after that the cluster tree is formed. Yunsha et al [3], lexical graph is generated where the edges are marked with association grade of the sentence. The nodes with higher degree for class name. Fahim A M et al.[4] discussed a efficient method for assigning data points to clusters in k-means clutering.

The main aim of this research paper is to proposed Genetic Algorithm Based Document Clustering(GABDC) method for documents in a text and then relating the same for classifying the sub topics in a collection of document. Related sentences are grouped together to form a fine cluster, and same sub topics. The proposed method selects the words from documents. It compares every word and identify the matching words. Initially sign of elements of the Eigen vectors of the similarity matrix used for the cluster.

III. METHODOLOGY

The proposed GABDC method extracts sentences from large documents. The stop words are removed from the documents and tokenized. The left behind set of words are stemmed as in Fig .4 and adjacency matrix A is used. The vertex degree is used for the diagonal matrix D to find out the similarity matrix C is equivalent to D-A. Eigen values taken from similarity

ISSN 2394-3777 (Print)
ISSN 2394-3785 (Online)
Available online at www.ijartet.com
International Journal of Advanced Research Trends in Engineering and Technology (IJARTET)
Vol. 3, Special Issue 20, April 2016

matrix, which is close to λ d = m / c is designated, whenever m is edges total number and sum of vertex degrees is c. the texts divided in to two cluster based on the elements sign in the Eigen vector equivalent to this Eigen cost. The heading stores the Meta data like the k is number of clusters, nk is number of elements in each cluster. The data portion of the document details is to selection, mutation and crossover and of unique documents. The applied fitness value to each and every document.

The proposed system architecture of GABDC is shown in Fig. 1.

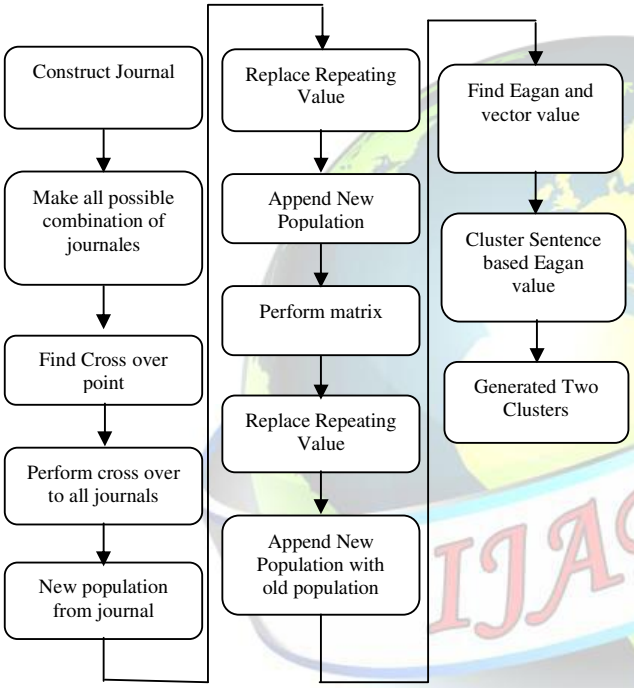

Fig. 1 System Architecture of GABDC

The fig.2 shows the pseudo code of the GABDC

Algorithm: GABDC

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 documents and make all the combinations of the cluster member part of the documents.
4. P_T has n! number of documents
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 documents.
7. Perform crossover of all documents.
8. Replace repeating elements with missing elements in all documents. This will create the new population Pc.
9. Add $P_0 + P_c$. Remove the duplicating documents.
10. Perform mutation of the documents by subtracting each element of the documents from the last element and take the absolute value.

cont.....

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 documents from P_{new}.
14. Find fitness of each documents
15. Remove documents with zero fitness value.
16. Do steps 6 to 15 , ten times
17. Select the documents with the highest fitness value which is the solution to clustering.

Fig. 2 pseudo code of GABDC

IV. Results and Discussion

The GABDC method is experimentally evaluated using MATLAB R2013a. The dataset collected from Reuters-21578 Text Categorization collection Dataset. This is a collection of documents that appeared on Reuters newswire in 1987. The documents were assembled and indexed with categories. That .is used to experiments this algorithm. This dataset is used to correct a variety of typographical and other errors in the categorization and formatting of the collection.

Fig.3, shows the sample document about three topics machine learning, genetic algorithm and cluster analysis from Reuter dataset.

```
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.
```

ISSN 2394-3777 (Print)
ISSN 2394-3785 (Online)
Available online at www.ijartet.com
International Journal of Advanced Research Trends in Engineering and Technology (IJARTET)
Vol. 3, Special Issue 20, April 2016

Fig. 3 Sample Document

Fig.4 Stemming refers to the process of reducing terms to their stems or root variants. For example generated-> generate; searched, searching -> search etc. Stemming reduces the computing time as different form of words is stemmed to form a single word.

mapr <6x15 cell>

	1	2	3	4	5	6	7
1	'Cluster'	'analysi'	'cluster'	'group'	'group'	'group'	'object'
2	'Cluster'	'analysi'	'common'	'data'	'explorator'	'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

Calculated Fitness values are using the mentioned formula in algorithm and step 14. Between each pair of sentences in a cluster is calculated. Obtained this value from the attributes in sentence term vector which represented in Fig.5

Jacard <6x51 double>

	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

Fig. 6 represents the basic operators of genetic algorithm. It is depends on performance of both crossover and mutation. The proposed work shows the outcome of the operations.

nwP6 <36x6 double>

	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 Documents after crossover, mutation.

```
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 Final Cluster output.

The Quality of the method has been evaluated using statistical measures precision. Precision means that a measurement gets similar results every single time it is used. It measures how well the tools are working not what the tools are measuring. Precision can be defined as description of random errors, a measure of statistical variability. The time taken to execute the algorithm has been calculated and compared.

Precision refers to the closeness of two or more measurements to each other. Accuracy refers to the closeness of a measured value to a standard or known value.

A. Precision vs number of clusters

The given cluster measure precision as the number of correct points that belong to the cluster over the total number of points in the cluster. The closer this the better the clustering algorithm for a fixed number of cluster.

Fig. 8 shows the comparison between K-means clustering method and GABDC method.

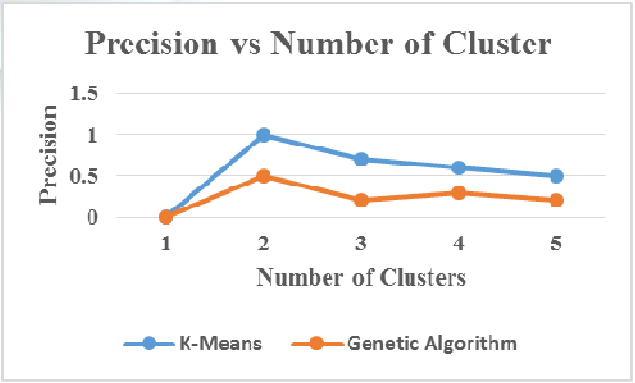

Fig .8 Comparison Precision values

ISSN 2394-3777 (Print)
ISSN 2394-3785 (Online)
Available online at www.ijartet.com
International Journal of Advanced Research Trends in Engineering and Technology (IJARTET)
Vol. 3, Special Issue 20, April 2016

Fig 8 shows that the precision reductions as the number of clusters increase and the genetic algorithm has a improved precision than k-means.

B. Time vs number of clusters

Fig. 9 shows between the number of clusters and time calculated for both genetic algorithm and k-means algorithm.

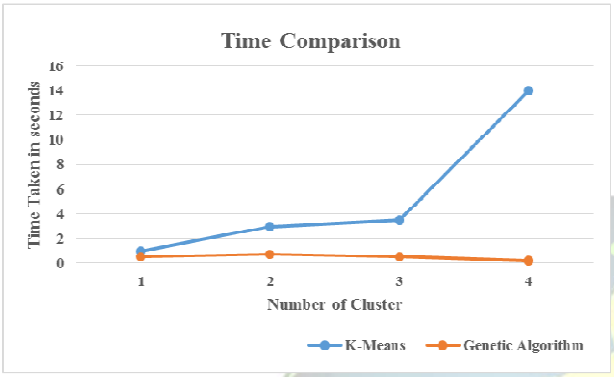

Fig .9 Time comparison.

Here from the graph K means algorithm takes slightly more time than genetic algorithm.

IV CONCLUSIONS

The result in this proposed work Genetic Algorithm Based Document Clustering(GABDC) gives more good performance and different methods give better results. Fitness function is a function which assigns fitness value to the individual. It is problem specific. And also mutation function changing random documents in an individual. Genetic algorithm is used to generate useful solutions to optimize and inspired by natural evolution. Compare to K-means algorithm, genetic algorithm improves the accuracy of documents. The time taken to execute the proposed method is less when compare other conventional methods. The overall performance of GABDC method is better than other methods.

REFERENCES

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

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

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

[4] Fahim A.M Salem A,M Torkey A and Ramadan M A,Än Efficient enhanced K-means clustering algorithm,"Journal of Zhejiang University,(10,7):1626-1633,2006.

[5] 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 .

[6] https://archive.ics.uci.edu/ml/datasets/Reuters-21578+Text+ Categorization+Collection

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