ISSN: 2278 – 1323
International Journal of Advanced Research in Computer Engineering & Technology (IJARCET)
Volume 5, Issue 12, December 2016

A Genetic approach to Classify Text Documents by Utilizing Term features

Shabana Bee[1], Mr. Sumit Gupta[2]
M.Tech Student[1], Assistant Professor[2]
Department of Computer Science.& Engg.
Lakshmi Narayan College of Excellence[12]

Abstract— As the text document are increasing day by day with the growing digital world. Researchers are working in this field from last few decades. In this paper a genetic algorithm is proposed on classify the text document in efficient manner. At present teachers learning algorithm is utilize for the classification which is known as genetic approach. Proposed classification approach classifies the data on the basis two phase learning where best probable solution is consider as the teacher phase while in second phase individual solution will learn from each other. In each phase new probable solution having high fitness value is preserve while low fitness value probable solution was discard. Experiment is done real as well as artificial dataset. Proposed work is compare with fire fly genetic approach and results shows that proposed work is better as compare to previous work on different evaluation parameters.

keywords— Document, genetic algorithum, feature extraction, text categorization, clustering

I. INTRODUCTION

As large amount of digital data has been collected on different servers for the various purposes. In case of document classification is the major problem in these days.This is because of high dimension availability in documents. [6]. As in classification similar type of objects are found in the dataset than group back in the field. Text document similarity is obtained by finding the similarity function. The difficulty of classification can be very helpful in the text province, where the matter to be classify can be of different dimension such as paragraphs, documents, sentences or terms.

For many research funding agencies, international journals, national journals, such as either government or private agencies, the selection of research project proposals is an important and challenging task, when large numbers of research proposals are collected by the organization. The Research Project Proposals Selection Process starts with the call for proposals, then from different research scholars, scientist, etc. from many institutes and organizations submit there research proposals. As there is single point of contact for researchers from different area so, group the proposals based on their similarity and assigned them to the experts for peer-review. The review results are examined and proposals are ranked based on their aggregation of experts result. So the simple steps of the Research Project Selection Process, these processes are very similar in all research funding agencies.[2]

For very large number of proposals received by the agencies need to be group the proposals for peer review. The department for selection process can assign the grouped proposals to the external reviewers for evaluation and rank them based on their aggregation. As they may not have adequate knowledge in all research discipline areas and the contents of many proposals were not fully understood when the proposals were grouped, there may be short of time for doing this so doing evaluation for whole in detail manually is tough. In current Methods, keywords are not representing the complete information about the content of the proposals and they are just the partial representation of the proposals. Hence, it's not sufficient to group the proposals on the basis of keywords. In Manual based grouping, sometimes the

department responsible for grouping may not have adequate knowledge regarding all the issues and areas of the research proposals. Therefore, an efficient and effective method is required to group the proposals efficiently based on their discipline areas by analyzing full text information of the proposals. So ontology is constructing for text-mining that will effectively used for this purpose.

II. Related Work

Wen Zhang et.al [1] has worked to developed an effective index for classifying the input documents into there respective category. Here work has given an comparison of various approaches such as TFIDF, LSI and multi word for text classification. It was observed from the result section of this paper that use of LSI was more effective then previous other methods. It is obtained that retrieval of the documents through LSI is more effective in case of English texts. So this work shows that LSI can produce the discriminative power for indexing as well.

Vishwanath Bijalwan et.al[2] has utilize the K-Nearest Neighbour method for clustering the documents into its category than further categorize and return the list in more relevant manner. Here results are compare with Naive Bayes and Term-Graph. It is obtained that proposed work has increase the accuracy of classification as compare to othe comparing methods. But KNN bring one drawback that include classification time, here time complexity of the work is quit high as compare to previous other work. Here use of AFOPT with KNN which is an hybrid approach works better as compare to individual one. Finally author made an information retrieval application using Vector Space Model to give the result of the query entered by the client by showing the relevant document.

Tanmay Basu et. al[3] As text document is of different dimension so classification is an tough task. Therefore, efficient method for feature selection is required to improve the performance of text classification. By the use of supervised term feature approach classification was got easy. Here comparison of proposed work is done with previous other approaches such as MI, CHI and IG. In this work as per the score obtained by the term a similarity rank was developed with the classifying categories. Here one more achievement was done by the work which has shown that proposed work achieved high classification accuracy even after removing the 90% unique contents.

Youngjoong Ko et. al[4] The main purpose of this paper is to improve text classification by efficiently applying class information to a term weighting scheme. Here classification was done in multiple category. Here comparison of proposed work was done with previous other methods and shows that by the use of use of term weight from the TFIDF gives better classification accuracy.

Aixin Sun et.ai [5] Here small text files are classified where number of class is independent and parametric independent. So utilization of this approach is in classifying the tweets, status, mimics, etc. It selects the representative words from a given short text as query words. After that it searches for a set of labeled text those best matches the query words. Here work was done on four independent approaches named as TF, TFIDF, TF.CLARITY and TF.IDF.CLARITY. Results obtained by classifying the real dataset and shows that classification accuracy is highest in case of TF.CLARITY.

III. Background

As the mining is utilize in different type of data analysis so for the same all need to increase the different technique in the required area. So contributing the text mining is done in this work by the proposed method for clustering the document or articles in the group without having any prior knowledge of the documents. In the propose work no need of any format for the input data such as speakers identification symbol or special character, here all process is done by utilizing the different combination of cluster center field.

a. Preprocessing

Preprocessing is a process used for conversion of document into feature vector. Just like text categorizations preprocessing also has controversy about its division [10] Text preprocessing is consisting of words which are responsible for lowering the performance of learning models. Data preprocessing reduces the size of the input text documents significantly. It involves activities like sentence boundary determination, natural language specific stop word elimination. Stop-words are functional words which occur frequently in the language of the text (for example a, the, an, of etc. in English language), so that they are not useful for classification.

This signifies that in maximum time words in corpus arises very few times in any training corpus. Those words which are arise very few times statically unimportant having low information gain. However the occurrence of any word in training in future document is very less.

The vector which contain the pre-processed data is use for collecting feature of that document. This is done by comparing the vector with vector KEY (collection of keywords) of the ontology of different area. So the refined vector will act as the feature vector for that research project/proposal.

So the list of words which are crossing the threshold are consider as the keywords or feature of that document.

[feature] = mini_threshold ([processed_text])

In this way feature vector is created from the document.

b. Generate Population

Here assume some cluster set that are the combination of different documents. This is generate by the random function which select fix number of document cluster for the centroid. This can be understand as let the number of centroid be Cn and number of documents are N then one of the possible solution is {C1, C2, …..Cn}. In the similar fashion other possible solutions are prepared which can be utilize for creating initial population represent by ST matrix.

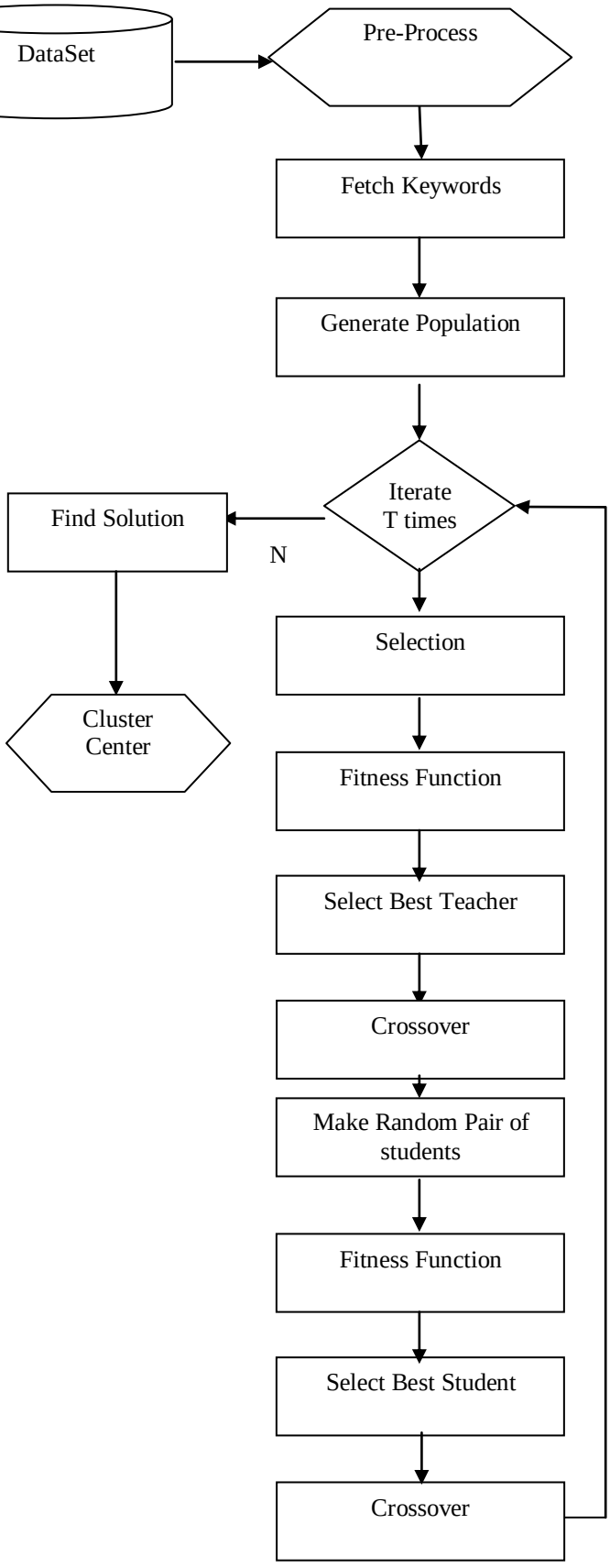

Fig. 1 Proposed work Block diagram.

ST[x] ←Random(N, Cn)

c. Teacher Phase:

For finding difference two chromosomes function are use first is Eludician Distance formula other is cosine similarity function

The Euclidean distance d between two solution X and Y is calculated by

d = [SUM((X-Y).^2)]^0.5

The Cosine distance d between two vectors X and Y is

d = [1 – (X*Y' / √(X*X')* (Y*Y'))]

Following Step will find distance between the selected population for finding the teacher in the population.

1. Loop x = 1:ST
2. Loop n = 1:N
3. D[n, x] = Dist(Ds[n], x) // Here Dist is a Euclidean function
4. endLoop
5. endLoop
6. S←Sum(D) // Sum matrix rowwise
7. [V I]←Sort(S) // Sort matrix in increasing order

So the matrix D contain all the values of the centriod distance from the document then find the minimum distance which will evaluate specify best possible solution.

Top possible solution after sorting will act as the teacher for other possible solutions. Now selected teacher will teach other possible solution by replacing fix number of centroid as present in teacher solution. By this all possible solution which act as student will learn from best solution which act as teacher.

Main motive of this step is to find best solution from the generated population. Here each possible solution is evaluated for finding the distance from each centroid document so that document closer to the centroid are cluster together. Then calculate the fitness value which give overall rank of the possible solution.

As teacher have quality to transfer its knowledge to its student in easy manner. Although it is possible that all student will not catch knowledge from the teacher. But an average of the class move forward by the teacher as it train other student to the best solution. So by using some random number it is possible that student get some values from the teacher. Here teacher work for training the student for his best level that is towards Here solution will update if and only if the new chromosome is better as compare to previous one.

This difference modifies the existing solution according to the following expression

Xnew,i = Xold, i + Replacing Cluster value

Where Xnew,i is the updated value of Xold,i. Accept Xnew,i if it gives better function value.

d. Student Phase

In this phase all possible solution after teacher phase are group for self learning from each other. This can be understand as let group contain two student then each student who is best as compare to other will teach other solution. Teaching is similar as done in teacher phase, here replacing fix number of centroid is done which is similar as in best student of the group.

1. For i = 1: Pn
2. Randomly select two learners Xi and Xj, where i is not equal to j
3. If f (Xi) < f (Xj)

4. Xnew, i = Xold, i + ri (Xi − Xj) (for a minimization problem)
5. Else
6. Xnew, i = Xold, i + ri (Xj − Xi)
7. End If
8. End For

Accept Xnew if it gives a better function value. Once student phase is over then check for the maximum iteration for the teaching if iteration not reach to the maximum value then GOTO step of teacher phase else stop learning and the best solution from the available population is consider as the final centroid of the work. Now documents are cluster as per centroid.

IV. Experiment And Result

a. Experimental Setup

All algorithms and utility measures were implemented using the MATLAB tool. The tests were performed on an 2.27 GHz Intel Core i3 machine, equipped with 4 GB of RAM, and running under Windows 7 Professional.

b. DataSet

For experiment work is evaluate on the standard dataset named as 20 books, where each book contain different number of text files of the same field. So for experiment only few books are consider as input. As experiment results are almost same for the books, so out of thousands of files present in dataset only few of them are consider.

Dataset Description	
Book	**Number of File**
Dataset1	60
Dataset2	50

Table 1 Dataset Description for proposed work evaluation.

c. Evaluation Parameter

Centroid Distance:

In order to evaluate results there are many parameter such as Euclidean distance of the chosen centroid and its cluster documents. Obtaining values can be put in the mention parameter formula to get comparison values.

Execution time

the work done on the important resource that may be server so execution time should be as less as possible. So this is a very important parameter to evaluate this work. This is to evaluate execution time time of the algorithm that is time taken by the proposed method for execution. Algorithm time is expect after the evaluation of the direct and indirect rules.

d. Results

Results of the proposed work is compare with the existing method FA (Fire Fly Algorithm) in [3]. Eludician, distance of the chosen centroid and its cluster documents are calculate on the basis of the mention formula explained in previous section.

Dataset	Distance Function	FA	TLBO
Dataset1	Euclidean	4.2949	2.8694
Dataset2	Euclidean	9.4926	3.9391

Table 2 Centroid Distance on the basis of Eludician distance methods.

Dataset	Distance Function	FA	TLBO
Dataset1	Cosine	3.0002	2.0002
Dataset2	Cosine	8	3

Table 3 Centroid Distance on the basis of Cosine distance methods.

Above table 2 and 3 shows that FA distance between its centroid and cluster document is high as compare to the proposed work of TLBO. It has been observed that proposed work centroid selection method is efficient as compare to the FA. Results are evaluate on two distance formula first is eludician and other is Cosine, in both the case values is TLBO values are less as compare to FA.

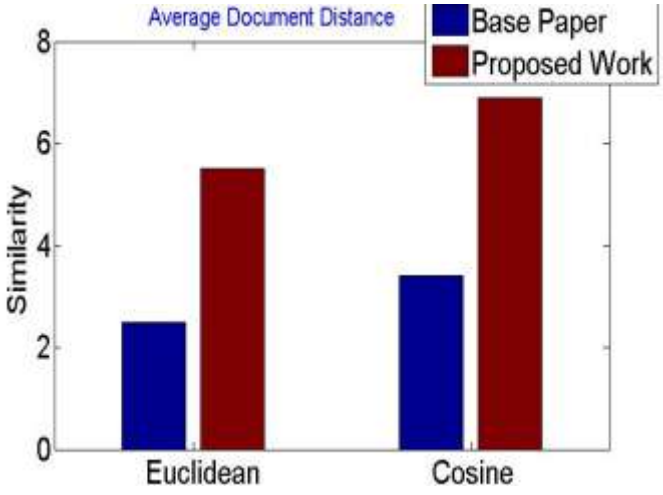

Fig. 2 Average distance between FA and TLBO Algorithm.

Above graph 2 shows that average of distance of all the dataset from the choose centroid of the TLBO algorithm is less as compare to the FA algorithm. So clustering done by the proposed work is much better as compare to the existing work of FA in all conditions.

Dataset	FA (sec)	TLBO (sec)
Dataset1	7.3722	4.56
Dataset2	4.7611	4.56

Table 4 Execution Time of TLBO and FA algorithm.

Above table 4 it has been observed that execution time for FA algorithm for clustering document is high as compare to the proposed work of TLBO. It has been observed that proposed work centroid selection method is efficient as compare to the FA.

VI. Conclusions

With the drastic increase of the digital text data on the servers, libraries it is important for researcher to work on it. Considering this fact work has focus on one of the issue of the document classification which is build by the different organization such as news, debate, online articles, etc. Here many researchers has already done lot of work but that is focus only on the content classification where in this work document are classify. In few work document classification are done on the basis of the background information, but this work overcome this dependency as well here it classify all the document without having prior knowledge. Results shows that using an correct iteration with fix number of centroid for classification proposed algorithm works better then previous FA one. As there is always work remaining in every because research is a never ending process, here one can implement similar thing for different other language.

References

[1] Wen Zhang, Taketoshi Yoshida, Xijin Tang. "A Comparative Study Of TF*IDF ,LSI And Multi Words For Text Classification",2011,Vol.1.

[2] Vishwanath Bijalwan, Vinay Kumar, Pinki Kumari, Jordan Pascual. "KNN Based Machine Learning Approach For Text And Document Mining", 2014,Vol.7,No.1,Pp.61- 70.

[3] Tanmay Basu, C. A. Murthy, "Effective Text Classification By A Supervised Feature Selection Approach",2008.

[4] Youngjoong Ko, "A Study Of Term Weighting Schemes Using Class Information For Text Classification", Aug 12- 16,2012.

[5] Aixin Sun, "Short Classification using very few words", 2012, ACM 978-1-4503-1475-5/12/08.

[6] Selma Ayşe Özel. Esra Saraç " Web Page Classification Using Firefly Optimization ", 978-1-4799-0661-1/13/$31.00 ©2013 IEEE.

[7] M. Nagy And M. Vargas-Vera, "Multiagent Ontology Mapping Ramework For The Semantic Web," IEEE Trans. Syst., Man, Cybern. A, Syst., Humans, Vol. 41, No. 4, Pp. 693–704, Jul. 2011.

[8] G. H. Lim, I. H. Suh, And H. Suh, "Ontology-Based Unified Robot Knowledge For Service Robots In Indoor Environments," IEEE Trans. Syst., Man, Cybern. A, Syst., Humans, Vol. 41, No. 3, Pp. 492–509, May 2011.

[9] Q. Liang, X. Wu, E. K. Park, T. M. Khoshgoftaar, And C. H. Chi, "Ontology-Based Business Process Customization For Composite Web Services," IEEE Trans. Syst., Man, Cybern. A, Syst., Humans, Vol. 41, No. 4, Pp. 717–729, Jul. 2011

[10] H. C. Yang, C. H. Lee, And D. W. Chen, "A Method For Multilingual Text Mining And Retrieval Using Growing Hierarchical Self-Organizing Maps," J. Inf. Sci., Vol. 35, No. 1, Pp. 3–23, Feb. 2009.

[11] Guansong Pang, Shengyi Jiang, " A Generalized Cluster Centroid Based Classifier For Text Categorization",2013.

.

.