INTERNATIONAL JOURNAL OF RESEARCH IN COMPUTER APPLICATIONS AND ROBOTICS
ISSN 2320-7345

A ROBUST DOCUMENT RETRIEVAL APPROACH WITH GENETIC APPROACH TO CLASSIFY TEXT DOCUMENTS

Shruti pathak, prof. Manish misra

Abstract— As the text document are increasing day by day with the growing digital world. It is required to read and classify the data on daily basis. Researchers are working in this field from last few decades. In this paper a genetic algorithm is proposed that classify the text document in efficient manner. Here teachers learning algorithm is utilize for the classification which is a genetic approach. Proposed classification approach classifies the data on the basis of terms features. Here two phase learning help in effective classification technique. After perfect classification of document retrieval of document as per text query is done. Results shows that proposed work is better as compare to previous work on different evaluation parameters.

Index Terms— Information Extraction, Text Analysis, Ontology, feature extraction, text categorization, clustering

I. INTRODUCTION

The problem of clustering has been studied widely in the database and statistics literature in the context of a wide variety of data mining tasks [50, 54]. The clustering problem is defined to be that of finding groups of similar objects in the data. The similarity between the mining text data objects is measured with the use of a similarity function. The problem of clustering can be very useful in the text domain, where the objects to be clusters can be of different granularities such as documents, paragraphs, sentences or terms.

For many research funding agencies, international journals, national journals, such as 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

Vishwanath Bijalwan et.al [1]:- In this paper author have first categorized the documents using KNN based machine learning and then return the most relevant documents. In this paper author conclude that KNN shows the maximum accuracy as compared to the Naive Bayes and Term-Graph. The disadvantage of KNN classifier is that its time complexity is high but gives a enhanced accuracy than others. In this paper the author rather than implementing the traditional Term-Graph used with AFOPT used Term-Graph with other methods. This hybrid shows a better result than the traditional combination. 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.

Wen Zhang et.al [2]:- The purpose of this paper is to study the effectiveness of different indexing methods in text classification. This paper has comparatively studied TF_IDF, LSI and multi-word for text representation. An experimental result has demonstrated that in text classification, LSI has performed very well than other methods in both document collections. Also, while retrieving English documents LSI showed the best performance. This outcome has shown that LSI has both favorable semantic and statistical quality and is different with the claim that LSI cannot produce discriminative power for indexing.

Tanmay Basu et.al [3]:- Text classification is a difficult task due to its high dimensionality of data. Therefore, efficient method for feature selection is required to improve the performance of text classification. This paper presents a new feature selection method for text classification using a supervised term selection approach. In this paper TS(term significance) a feature selection technique is compared with CHI,IG & MI. The proposed approach derives a similarity score between a term and a class and then ranks the terms according to their scores over all the classes. The experimental results show that the proposed TS can produce better classification accuracy even after removing 90% unique terms.

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. The author purposed a new scheme for multi class text classification. Then it was compared to the TF-IDF and previous methods. As a result the proposed scheme utilized class information for term weighting for text classification and performed consistently on the data sets and KNN and SVM classifiers.

Aixin Sun et.ai [5]:- In this paper the author purposed a simple, scalable and non-parametric approach for short text classification. This approach mimics human classification process for a piece of short text like tweets, status updates, and comments. 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. The author have used four approaches and are evaluated to select the query words: TF, TF.IDF, TF.CLARITY and TF.IDF.CLARITY. Experimental result shows that TF.CLARITY performs effectively when three or more words are used in a query whereas TF.IDF.CLARITY performs well when one word is used in a query.

III. Proposed Methodology

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.

Preprocessing

Preprocessing is a process used for conversion of document into feature vector. Just like text categorizations the preprocessing also has controversy about its division. The preprocessing is divided into two parts – text preprocessing and document indexing [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 and stemming. 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. When categorization is done pruning mostly produces feature space of small size.

Fetch KeyWords

The vector which contains 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.

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.

ST[x] ←Random(N, Cn)

Teacher Phase:
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 populations for finding the teacher in the population

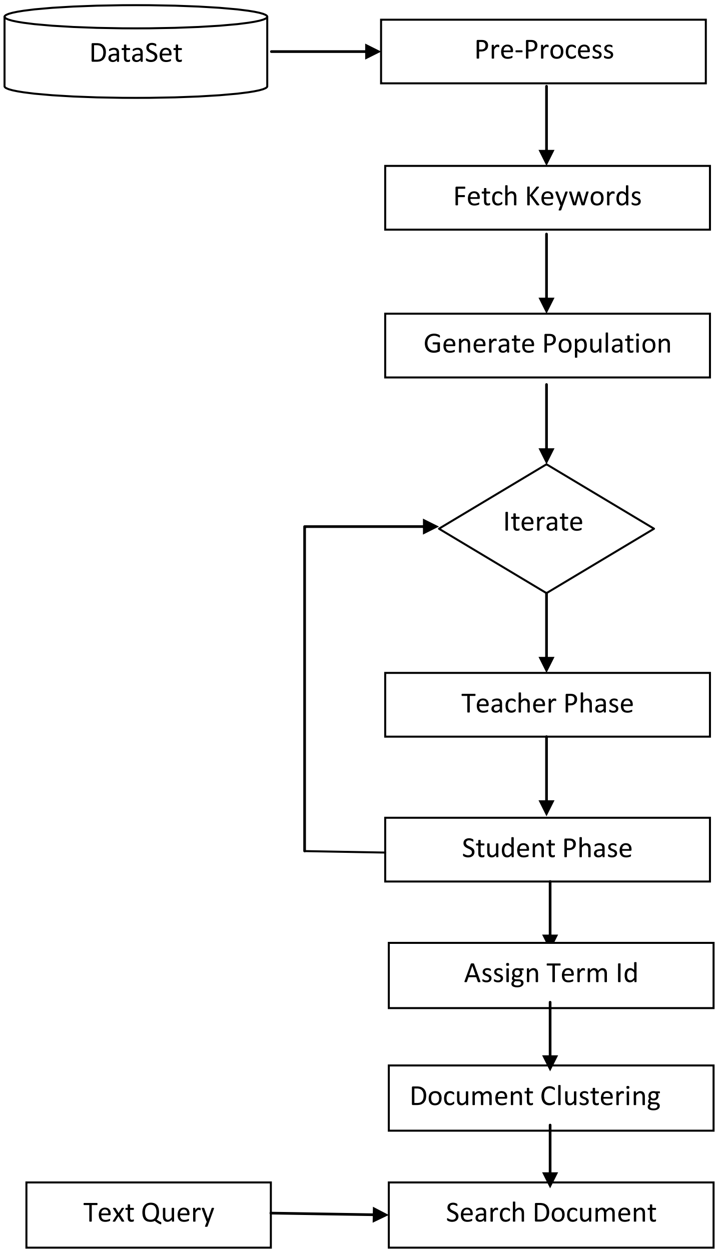

Fig. 1 Proposed work Block diagram.

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

```
Loop x = 1:ST
    Loop n = 1:N
        D[n, x] = Dist(Ds[n], x) // Here Dist is a Euclidean function
    endLoop
endLoop
S←Sum(D)     // Sum matrix row wise
[V I]←Sort(S)       // Sort matrix in increasing order
```

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

S←Sum (D)   // Sum matrix row wise
[V I]←Sort(S)   // Sort matrix in increasing order

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 gives overall rank of the possible solution.

A good teacher is one who brings his or her learners up to his or her level in terms of knowledge. But in practice this is not possible and a teacher can only move the mean of a class up to some extent depending on the capability of the class. This follows a random process depending on many factors. So Let teacher is T =[C1 ,C20, C5 ,C11] used student is ,S =[C7,C15,C9,C1].Now

Teach T will teach student S by randomly replacing one cluster value in from T*to S. Here if random position is 2 then C20 is place in S will this replace C15 cluster value as S So new cluster to [C7,C20 ,C5, C11]

This difference modifies the existing solution according to the following expression

$$X_{new,i} = X_{old,\ i} + \text{Replacing cluster value}$$

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

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.

```
For i = 1: Pn
    Randomly select two learners Xi and Xj, where i is not equal to j
    If f (Xi) < f (Xj)
        Xnew, i = Xold, i + ri (Xi − Xj) (for a minimization problem)
    Else
        Xnew, i = Xold, i + ri (Xj − Xi)
    End If
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.

Assign Term-Id

In this step keywords obtained from the features of the document are need to be inserted into the neural network for classification but as we know that text words cannot be inserted into the neural network. So the representative of those words is required. As each keyword is a set of ASCII value for example keyword "ABCD" ASCII set is [65 66 67 68]. Now each ASCII number is replace by its binary number as 65={ 1000001}, 66={ 1000010}, 67={ 1000011}, 68={ 1000100}. So in this, we will replace ABCD by its binary number that is {1000001100001010000111000100}.

Search Document

As each word contains a different number of characters so a set of 100 bit is taken as input in the neural network, where the default value is zero in the vector.

In this step as per the keywords (terms) from the user text query have their own term id while number are same as present in the dataset. Due to this term id privacy of the user query is increases. Now all term-id that are present in the text query act as key for the selecting the cluster where each document set from the matched cluster are index as per the text query term-id.

IV. Experiment And Result

In order to implement above algorithm for document retrieval MATLAB 2012a tool was used. Here same work can be implementing on other programming language as well. But as some of the function was inbuilt in the tool which help researcher to focus on the work. Experiment was done on real as well as on artificial dataset. Here different set of dataset was use for retrieving documents.

Evaluation Parameter

As various techniques evolve different steps of working for classifying document into appropriate category. So it is highly required that proposed techniques or existing work need to be compare on same dataset. So following are some of the evaluation formula shown in equation number 4, 5, 6and 7 which help to judge the classification techniques ranking.

Precision = (True_positive / (False_positive+ True_positive))

True positive Rate = (True_positive /(False_negative+ True_positive))

F-Measure = (2xPrecisionxRecall/ (Recall + Precision))

A. Results

Comparison of Recall		
Query	Proposed Work	Previous work[9]
Q1	0.545455	0.5
Q2	0.5	0.363636
Q3	0.545455	0.454545

Table. 1. Comparison of recall value with previous work [9].

From above table 1 it is obtained that proposed work recall value is higher than previous work on different queries. As query set has good quality keywords results of proposed work is also high.

Comparison of precision values		
Query	Proposed Work	Previous work[9]
Q1	0.857143	0.714286
Q2	0.7143	0.571429
Q3	0.857143	0.714286

Table. 2. Comparison of precision value with previous work [9].

From above table 2 it is obtained that proposed work precision value is higher than previous work on different queries. As query set has good quality keywords results of proposed work is also high.

Comparison of F-Measure values		
Query	Proposed Work	Previous work[9]
Q1	0.666667	0.588235
Q2	0.588235	0.444444
Q3	0.666667	0.555556

Table. 3. Comparison of F-Measure value with previous work [9].

From above table 3 it is obtained that proposed work F-Measure value is higher than previous work on different queries. As query set has good quality keywords results of proposed work is also high.

Comparison of Execution time in second		
Query	Proposed Work	Previous work[9]
Q1	0.011595	1.79145
Q2	0.00271972	2.02409
Q3	0.00721962	2.10903

Table. 3. Comparison of execution time in second with previous work [9].

From above table 3 it is obtained that proposed work execution time in second is lower than previous work on different queries. As binary value make easy comparison of the words. So overall execution time got reduced.

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 and retrieval with privacy of the content 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 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] 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.

[2] Wen Zhang, Taketoshi Yoshida, Xijin Tang. “A Comparative Study of TF*IDF, LSI and multi words for text classification”,2011,Vol.1.

[3] Tanmay Basu, C. A. Murthy, “Effective Text Classification by a Supervised Feature Selection Approach”,2008.

[4] Guansong Pang, Shengyi Jiang, “ A Generalized Cluster Centroid based classifier for text categorization”,2013.

[5] Youngjoong Ko, “A Study of Term Weighting Schemes Using Class Information for Text Classification”, Aug 12- 16,2012.

[6] M. Nagy and M. Vargas-Vera, “Multiagent ontology mapping Framework for the semantic web,” IEEE Trans. Syst., Man, Cybern. A, Syst., Humans, vol. 41, no. 4, pp. 693–704, Jul. 2011.

[7] 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.

[8] S.Ramasundaram, “Ngramssa Algorithm For Text Categorization”, International Journal Of Information Technology & Computer Science (Ijitcs), Volume 13, Issue No : 1, Pp.36-44, 2014.

[9] Chi Chen, Member, Xiaojie Zhu, Peisong Shen, Jiankun Hu, Song Guo, Zahir Tari, Albert Y. Zomaya. “An Efficient Privacy-Preserving Ranked Keyword Search Method”. Ieee Transactions On Parallel And Distributed Systems, Vol. 27, No. 4, April 2016

[10] Wenhai Sun, Bing Wang, Ning Cao, Ming Li, Wenjing Lou, Y. Thomas Hou And Hui Li . “Verifiable Privacy-Preserving Multi-Keyword Text Search In The Cloud Supporting Similarity-Based Ranking”. Ieee Transactions On Parallel And Distributed Systems, Vol. 25, No. 11, November 2014.