Available Online at www.ijcsmc.com

International Journal of Computer Science and Mobile Computing

A Monthly Journal of Computer Science and Information Technology

ISSN 2320–088X
IMPACT FACTOR: 6.017

IJCSMC, Vol. 6, Issue. 3, March 2017, pg.236 – 241

An Efficient Text Classification Scheme Using Clustering

Kiran Khairnar[1], Swapnil Pagare[2], Prathamesh Choudhari[3], Sankalp Redgaonkar[4]

[1]Computer Engineering & Savitribai Phule Pune University, India

Kirankhirnar1993@gmail.com; Swapnil.pagare1000@gmail.com; prathamesh451@gmail.com; sankalpr007@gmail.com

Abstract— We are dealing with large amount of data today like text, image, and spatial form. So there is great significance of text mining process now a days. There exist many algorithms for text classification but they are having several drawbacks like accuracy, time consumption etc. For overcoming these we are using dimension reduction technique like SMTP, Auto encoder, PCA etc. In our technique we are creating several clusters and similarity measures are used for calculating similarity of new input document and created clusters. Clustering makes use of labelled texts to capture images of text clusters and unlabeled text to adopt its centroids. While the similarity is calculated, the clusters that matches the best to the input documents will get that document in it. User can manually change document location and put it any cluster he wants and system will
Self-learn the user instruction and work accordingly from next input document.

Keywords— Clustering, Similarity Measures, Data Mining, Classification, Semi-supervised.

I. INTRODUCTION

Text mining is similar to data mining, here the data is in the form of text. We use this when we need to get data from set of text documents. Cluster are group of similar item sets. We can make clusters on the basis of distance between nodes or on the basis of similarity measures. This process of making cluster is known as clustering. These things when brought together can make a new system which we term as ”Text Classification Using Clustering”. Document is a template.

But the thing that leads to the failure of system or result in drawback is calculating inaccurate similarity value, less efficiency, Time taken for complete processing, No manual changes in clusters by user.
Considering these drawbacks we will be designing a system that will overcome these drawbacks. In our system we will be using more efficient and accurate function known as SMTP. Also we will be preprocessing the Document which we will take as input. Processing will consist of Extraction of document and Stop word removal.

Stop word removal will reduce the time taken for further processing and give better time complexity. This can be done by Dimension Reduction Technique, Document term matrix, SMTP based similarity measures, Matching SMTP and forming cluster from input query document.

Hence, the project idea is to achieve better efficiency and time complexity for text classification using clustering. We will also provide user to change the location of documents in cluster with machine learning function.

II. RELATED WORKS

An easy way to comply with the conference paper formatting requirements is to use this document as a template and simply type your text into it. Text clustering and text classification are the two important text mining task. The clustering can be used to aid the text classification either as an alternative approach to term selection for dimensionality reduction or as a technique to enhance the training set. In the second approach clustering is used to discover the kind of structure in training examples. The model for classification is constructed using the extracted clusters. [1]

TESC (Text classification using Semi-supervised Clustering) is an approach for text classification using clustering proposed by Zhang et al. In this work the task of constructing classification model is done using a semi-supervised clustering and this model is then further used to assign the correct classlabels to the new document of the domain. [2]

A tremendous amount of side information is available and can be used for the text mining to improve the performance. Side information means the extra information or the metadata provided in the documents. This includes the document provenance information, locations, web logs, hyperlinks etc. [3]

Charuetal, used these side in formation to improve the performance of classification and clustering. But there are problems with the proper handling of side information and the noisy side information may affect the performance of the mining. [4]

Liu et al. generalize a boosting framework for improving the supervised learning algorithm with unlabelled data known as semi-boost. It improves the classification accuracy iteratively. Similarity measures play a significant role in classification and clustering. So proper selection of the suitable measure is an important step in text mining. Similarity measure is a real valued function that measures the similarity between two objects. SMTP is an efficient similarity for text classification and clustering and it satisfies all the desirable properties for a good similarity measure. [5]

III. MOTIVATION OF PROJECT

Document clustering has been used in many different areas of text mining and information retrieval. Initially it was used for improving the precision and recall in information retrieval systems and finding nearest neighbors of a document. Later it has also been used for organizing the results returned by a search engine and generating hierarchical clusters of documents.

Initially there are the techniques such as Euclidean distance measure and Cosine similarity measure clustering methods on the data and found that the results were not very satisfactory and the main reason for this was the noise in the text document. This provided the motivation for trying a pre-processing of the documents to remove the noise and outliers. System will be used to apply similarity measure for text processing (SMTP) and Dimensionality Reduction methods to get much better results.

There will be completely different approach by using SMTP and Dimensionality Measure for efficient classification of the document using clustering.

IV. SYSTEM DESCRIPTION

***EXTRACTION*:** Extraction is the process of extracting the data from the source. Extraction is done through tokenization of the file. Tokenization is the process of breaking a stream of text into words, phrases, symbols or other meaningful elements called tokens. This extracted data is then used to store in data warehouse after further processing on that.

***STOP WORD REMOVAL**:* Stop words are basically a set of commonly used words in any language, not just English. The reason why stop words are critical to many applications is that, if we remove the words that are very commonly used in a given language, we can focus on the important words instead. Stop words are language specific words which carry no information. The most commonly used stop words in English are e.g.: is, a, the etc.

***STEMMING**:* Stemming is the process for reducing inflected words to their word stem (base form).”Stemming attempts to remove the differences between inflected forms of a word, in order to reduce each word to its root form. For instance foxes may be reduced to the root fox, to remove the difference between singular and plural in the same way that we removed the difference between lowercase and uppercase.

VECTOR FORMATION: Vector formation is the process of converting the text contents into numeric format for further processing. For that we either use the term frequency (tf), inverse document frequency (idf), term frequency-inverse document frequency (tf-idf). In our experiment tf-idf is used to generate the document vectors and finally a term document matrix is formed.

Document classification or document categorization is a problem in library science, information science and computer science. The task is to assign a document to one or more classes or categories. The documents to be classified may be texts, images, music, etc. Each kind of document possesses its special classification problems.

When not otherwise specified, text classification is implied. The system contains a dataset which contains a various text files on which system perform the preprocessing steps: Extraction, stop word removal, stemming and vector formation.

System is using the porter stemmer algorithm for stemming the document. Document term frequency is then calculated by the term frequency inverse term frequency which is used to generate the document vectors and finally a term document matrix is formed. Documents are clustered by according to their similarity features using the SMTP algorithm. SMTP gives the better and efficient results than the other methods such as cosine, Euclidean distance formula and dice coefficient.

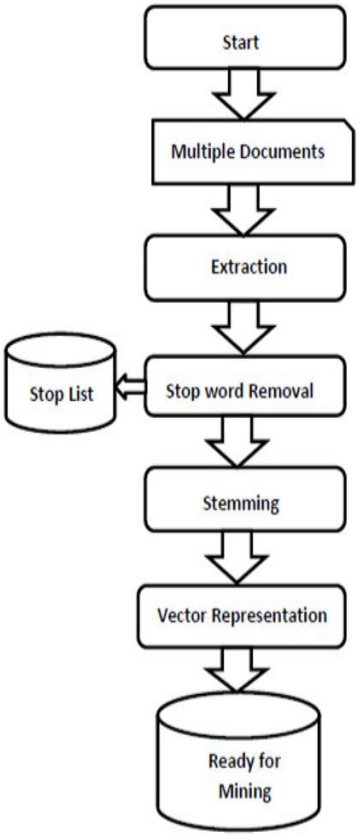

Fig 1. The Preprocessing Steps.

V. EXPERIMENTAL SETUP

The main aim of this work is to improve the performance of classification not the clustering. The data set that system will for the experiments is Reuters-21578. We took 200 documents for our experiments and that consist of labelled documents. Labelled documents are those documents that got class information. The classes that we are using in our experiment are faculty, student, project, and course. The model construction phase is one of the most important steps in classification. In order to obtain better classification accuracy good model construction methods should be adopted. The use of semi-supervised clustering for the model creation is a better approach. Most of the classification algorithms need labeled documents in the training phase for the classification model generation. The clusters formed and the corresponding number of documents in each cluster. Following is the feature vector formation example

Doc1: 1) Banana is sweet
2) Banana is good to health
Doc2: 1) Apple is better than banana
2) Apple and apple pie are good to health
W= [apple, banana, good, health, pie, sour, sweet]
D1= [0, 2, 1, 1, 0, 0, 1]
D2= [3, 1, 1, 1, 1, 0, 0]
m=7, λ=1, σ=2

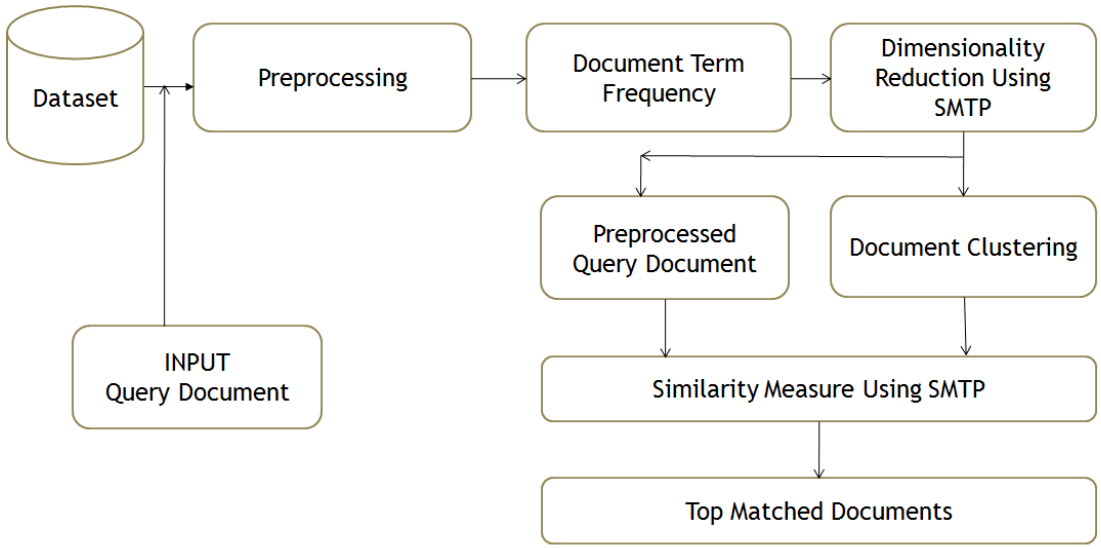

Figure Block Diagram

In our system we have put all the similarity measure formulae to calculate the similarity measure of the documents and to put in respected clusters. System contain Cosine Similarities, Extended Jaccard Coefficient similarities, Euclidean Distance Calculation, Dice Similarity Formula and important Similarity Measure For Text Processing(SMTP) Formula. System have the comparison graph for the SMTP, Cosine, and Euclidean Formulae which is drawn from the training dataset example. Also the user have access to change the cluster wherever he wants and then next time user can input similar document to previous document which is manually changed in system. It will put that document in user selected cluster automatically. System have the characteristics of automatic learning that is additional extension work to system.
As in graph you can see the SMTP is efficient technique than any other techniques.

VI. ALGORITHM

INPUT:
Documents in category C_i, (d_{i1},d_{i2},.....d_{in}) the clusters
S_i=i=1.....j when using SMTP algorithm and the test sample document X
OUTPUT:
Document X_s Category is C_i

Step 1:- Document X and all training sample are preprocessed stop-word removal, stemming done and the corresponding feature vector save in text file.
Step 2:- Normalization the columns of feature vector F by dividing highest word frequency to get the term frequency using formula:
Step 3:- Calculate weight of term to determine the different significance
$W_{ij} = TF_{ij}$ - Document term matrix
Step 4:- Select W_{ij} weights above a certain threshold value φ and also remove zero value. This term having less significance.
Step 5:- For every document distance is calculated using SMTP formula with all other document
for (i=0;i¡no. of document)
for(j=i+1;j¡no. of document)
Similarity=SMTP(i, j)
where i and j are the feature vector of document
Step 6:- Development of cluster is initiated by forming document -document group
Step 7:- Using a novel k-means algorithm with SMTP k no. cluster are constructed.
Step 8:- Performance of clustering is calculated using different similarity formulae e.g.: SMTP, cosine, Sine
Step 9:- Judge document x to be the category which has largest SMTP(X, C_j) C_j is a cluster centroid.
Step 10:- If X assign wrong cluster then take user choice cluster C_i and assign X to C_j
Step 11:- Update cluster centroid C_i and repeat step 9.

$$F(d_1, d_2) = \frac{\sum_{j=1}^{m} N_*(d_{1j}, d_{2j})}{\sum_{j=1}^{m} N_\cup(d_{1j}, d_{2j})} \quad (1)$$

Where

$$N_*(d_{1j}, d_{2j}) = \begin{cases} 0.5(1+\exp\left\{-\left(\frac{d_{1j}-d_{2j}}{\sigma_j}\right)^2\right\}, \\ \quad if\ d_{1j}d_{2j} > 0 \\ 0,\ if\ d_{1j} = 0\ and\ d_{2j} = 0 \\ -\lambda,\ otherwise, \end{cases} \quad (2)$$

$$N_\cup(d_{1j}, d_{2j}) = \begin{cases} 0,\ if\ d_{1j} = 0\ and\ d_{2j} = 0 \\ 1,\ otherwise. \end{cases} \quad (3)$$

Then SMTP for two documents is given by:

$$S_{SMTP}(d_1, d_2) = \frac{F(d_1, d_2) + \lambda}{1 + \lambda} \quad (4)$$

Fig: SMTP Formula

VII. CONCLUSION

It conclude that there is great importance of text mining process as we are dealing with large amount of data like text, image and spatial form so we are calculating similarity measure between the documents. Before the similarity measure, the document should go through various phases of data preprocessing such as extraction, stop word removal, stemming and vector representation.

Different similarity measures are applied in classification process and the results are produced from these techniques is not efficient so system is going to apply the SMTP based similarity measure for better results. It is efficient and time saving technique used for better results. After this future enhancement dimensionality technique is used and the dimension of term document matrix can be reduced. so that better execution time can be achieved and also the number of documents handled easily.

The applications like recommendation of news articles in case of news portals can be effective for this algorithm. Finally it conclude that through many algorithms have been proposed for clustering but it is still an open problem and looking at the rate at which the web is growing. So SMTP based similarity measure technique for clustering has good efficiency for text mining.

References

[1] Anisha Mariam Thomasa , Resmipriya M "An Efficient Text Classification Scheme Using Clustering", International Conference on Emerging Trends in Engineering, Science and Technology (ICETEST - 2015)
[2] Yung-Shen Lin, Jung-Yi Jiang, and Shie-Jue Lee, `A Similarity Measure for Text Classification and Clustering, IEEE Transactions On Knowledge And Data Engineering, Vol. 26, No. 7, July 2014.
[3] Charu C. Aggarwal, Yuchen Zhao, and Philip S. Yu, `On the Use of Side Information for Mining Text Data', IEEE Transactions on Knowledge and Data Engineering, Vol. 26, No. 6, June 2014.
[4] Pavan Kumar Mallapragada, Rong Jin, Member, Anil K. Jain, Fellow, and Yi Liu,`SemiBoost: Boosting for Semi-supervised Learning', IEEE Transctions On Pattern Analysis And Machine Intelligence, Vol. 31, No. 11, November 2009.
[5] Hung Chim and Xiaotie Deng, `Efficient Phrase-Based Document Similarity for Clustering, IEEE Transactions On Knowledge And Data Engineering, Vol. 20, No. 9, September 2008.
[6] Joris Dhondt, Joris Vertommen, Paul-Armand Verhaegen, Dirk Cattrysse, Joost R. Duon, `Pairwise-Adaptive Dissimilarity Measure for Document Clustering, Inf. Sci., vol. 180, no. 12, pp. 23412358, 2010.
[7] J. A. Aslam and M. Frost, "An information-theoretic measure for document similarity," in Proc. 26th SIGIR, Toronto, ON, Canada, 2003, pp. 449–450.
[8] J. Kogan, C. Nicholas, and V. Volkovich, "Text mining with information-theoretic clustering," Computer Sci. Eng., vol. 5, no. 6, pp. 52–59, 2003.
[9] T. Zhang, Y. Y. Tang, B. Fang, and Y. Xiang, "Document clustering in correlation similarity measure space," IEEE Trans. Knowl. Data Eng., vol. 24, no. 6, pp. 1002–1013, Jun. 2012.
[10] Wen Zhang, Xijin Tang, Taketoshi Yoshida, `TESC: An approach to Text classification using Semi-supervised Clustering', Knowledge-Based Systems, November 2014.

