A CLUSTERING TECHNIQUE FOR EMAIL CONTENT MINING

Deepa Patil and Yashwant Dongre

Department of Computer Engineering, VIIT, Pune, India

ABSTRACT

In today's world of internet, with whole lot of e-documents such, as html pages, digital libraries etc. occupying considerable amount of cyber space, organizing these documents has become a practical need. Clustering is an important technique that organizes large number of objects into smaller coherent groups. This helps in efficient and effective use of these documents for information retrieval and other NLP tasks. Email is one of the most frequently used e-document by individual or organization. Email categorization is one of the major tasks of email mining. Categorizing emails into different groups help easy retrieval and maintenance. Like other e-documents, emails can also be classified using clustering algorithms. In this paper a similarity measure called Similarity Measure for Text Processing is suggested for email clustering. The suggested similarity measure takes into account three situations: feature appears in both emails, feature appears in only one email and feature appears in none of the emails. The potency of suggested similarity measure is analyzed on Enron email data set to categorize emails. The outcome indicates that the efficiency acquired by the suggested similarity measure is better than that acquired by other measures.

KEYWORDS

Similarity Measure, Clustering Algorithm, Document Clustering, Email Mining

1. INTRODUCTION

We are living in cyber world dumped with whole lot of information. Efficient and accurate retrieval of the information is important for survival of many web portals. Text processing is an important aspect of information retrieval, data mining and web search. It is equally important to cluster the emails into different groups, so as to retrieve the similar emails i.e. email containing search string or key words. We can further group the mails based on time stamp for more easy retrieval.

As the amount of digital documents has been increasing dramatically over the years with the growth of internet, information management, search and retrieval have become an important concern. Similarity measure plays an important role in text classification and clustering algorithm. Many a times, a bag of words model [1], [2], [3] is used in information retrieval or text processing task, wherein a document is modelled as a collection of unique words that it has and their frequencies. The order is completely ignored. The importance of a word in the document can be decided based on term frequency (number of times a particular a word appears in the document), relative term frequency (it is ratio between term frequency and total number of occurrences of all terms in the document set) or tf-idf(it is combination of term frequency and inverse document frequency)[4]. All documents get converted into a matrix, where each word adds a new dimension, being row of matrix and each document is represented as a column vector. This implies that each entry in the matrix gives the frequency of world occurring in a particular document. It is easy to see that the matrix would be sparse. Higher the frequency of each word, it is more descriptive of the document.

DOI:10.5121/ijcsit.2015.7306

Clustering is a technique that organizes large number of objects into smaller coherent groups. Clustering aims at grouping similar documents in one group and separate this group as much as possible from the one which contains information on entirely different topics. Clustering algorithms [4] requires a metric to quantify how similar or different two given documents are. This difference is often measured by some distance measures. These measures are called as Similarity Measures. Clustering requires definition of a distance measure which assigns a numeric value to the extent of difference between two documents and which the clustering algorithm uses for making different groups of a given dataset. Similarity measure plays an important role in text classification and clustering.

A lot of similarity measures are in existence to compute similarity between two documents. Euclidian distance [5] is one of the popular similarity metric taken from Euclidian geometry field. Cosine similarity [4] is a measure which takes cosine of the angle between two vectors. The Jaccard coefficient [6] is a statistic used for comparing the similarity of two document sets and is defined as size of intersection divided by size of union on sample data sets. An information-theoretic measure for document similarity called IT_Sim [7], [8] is a phrase-based measure which computes the similarity based on Suffix Tree Document Model. Pairwise-adaptive similarity [9] is a measure which selects a number of features dynamically out of document d1 and document d2. In [7], [10] Hamming distance is used; hamming distance between two document vectors is number of positions where the corresponding symbols differ. The Kullback-Leibler divergence [11] is a non-symmetric measure of difference between probability distributions associated with two vectors.

Many a times, it is essential to search email contents to retrieve mails containing similar contents or key world .It is essential equally to search email body or content to put them into different groups or to search and bring up the emails that contains the search string or search keywords. This feature will be very useful in business domain. For example, if an employee in an organization wants to retrieve emails containing information about recent sells occurred in a particular area, he/she can specify the search string and search options accordingly, and fetch the emails containing required information.

My work focuses on implementing k-means clustering algorithm [12], [13] along with similarity measure (SMTP) Similarity Measure for Text Processing [14] on email data set to categorize emails into different groups. The main purpose of this work is to test effectiveness of SMTP used with k- means clustering algorithm for email clustering.

SMTP has many advantages. SMTP considers presence or absence of features than difference between two values associated with present feature. It also considers that similarity degree should increase when difference between two non-zero values of a specific feature decreases. It also takes into account that similarity degree should decrease when the number of presence-absence features increases. SMTP takes into consideration one more important aspect that two documents are least similar to each other if none of the features have non-zero values in both documents. SMTP is symmetric similarity measure. The last and most important fact is that it considers standard deviation of feature taken into count for its contribution to similarity between two documents.

The rest of the paper is organized as follows. Related work is discussed in short in section 2 .Proposed system is described in section 3. Experimental results are presented in section 4. Finally conclusion is given in section 5.

2. Related Work

Clustering is one of the data mining methods which is used for the purpose of email mining. Clustering is used for grouping emails for the purpose easy management .Commonly used clustering algorithms for email grouping are Hierarchical clustering [15] and k-means clustering algorithm [12],[13]. Closeness of any two emails can be determined by any distance measures such as Euclidian distance, Cosine similarity, Pairwise adaptive similarity, Jaccard coefficient, Dice coefficient etc.

The Euclidian distance measure [5] is defined as root of square differences between respective coordinates of d1 and d2 i.e.

$$d(p,q) = \sqrt{(\mathrm{p1}-\mathrm{q1})^2 + (\mathrm{p1}-\mathrm{q1})^2 + \cdots + (\mathrm{pn}-\mathrm{qn})^2}$$
$$= \sqrt{\sum_{i=0}^{n} (\mathrm{pi}-\mathrm{qi})^2} \tag{1}$$

Cosine similarity [4] measures the cosine of the angle between d1 and d2 as

$$similarity = \cos(\theta) = \frac{\mathrm{A.B}}{||\mathrm{A}||\ ||\mathrm{B}||} = \frac{\sum_{i=0}^{n} \mathrm{Ai}*\mathrm{Bi}}{\sqrt{\sum_{i=0}^{n} (\mathrm{Ai})^2} * \sqrt{\sum_{i=0}^{n} (\mathrm{Bi})^2}} \tag{2}$$

The resulting similarity ranges from −1 means exactly opposite, to 1 means exactly the same, with 0 usually indicating independence, and in-between values indicating intermediate similarity or dissimilarity.

The formula for Jaccard coefficient [6] for data processing is:

S_J = a/ (a + b + c), where (3)

S_J = Jaccard similarity coefficient
a = number of terms common to (shared by) both documents
b = number of terms unique to the first document
c = number of terms unique to the second document
Jaccard coefficient uses presence/absence data.

Dice coefficient is similar to Jaccard's index
Dice coefficient also uses presence/absence data and is given as:-
S_S = 2a/ (2a + b + c), where (4)

S_S = Dice similarity coefficient
a = number of terms common to (shared by) both documents
b = number of terms unique to the first document
c = number of terms unique to the second document

IT_Sim [7], [8] is a phrase-based measure to compute the similarity based on Suffix Tree Document Model. Pairwise-adaptive similarity [9] dynamically selects a number of features out of d1 and d2. Hamming distance [7], [10] between two vectors is the number of positions at which the corresponding symbols are different. The Kullback-Leibler divergence [11] is a non-symmetric measure of difference between probability distributions associated with two vectors.

Studies have shown that above mentioned similarity measures does not give optimal results for text classification. So a measure known as (SMTP) Similarity Measure for Text classification [14] is used for email categorization along with k-means clustering algorithm.

3. PROPOSED SYSTEM

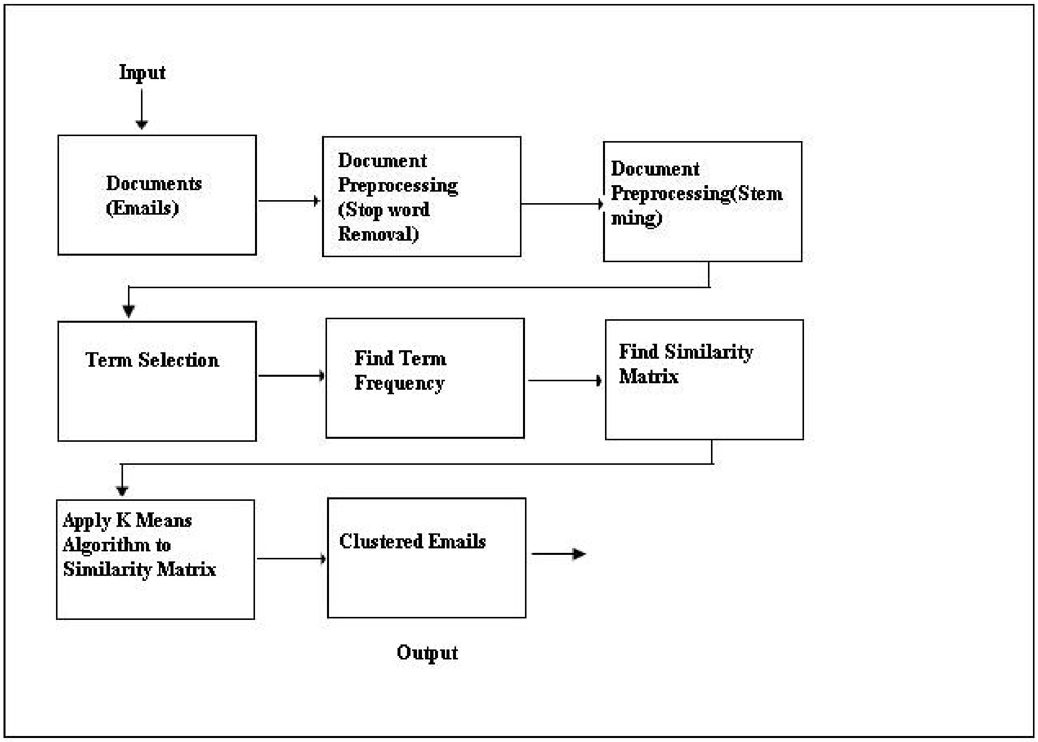

Input
Documents (Emails)
Document Preprocessing (Stop word Removal)
Document Preprocessing(Stem ming)
Term Selection
Find Term Frequency
Find Similarity Matrix
Apply K Means Algorithm to Similarity Matrix
Clustered Emails
Output

Figure 1. Proposed System

Email clustering will be carried on, on email data set.

Step 1:- Email data is pre-processed. Following are steps for pre-processing-

1. Removal of stopwords
2. Stemming process

Step 2:- Keywords are identified.
Step 3:- Term frequency is calculated
Step 4:-Similarity calculation uses distance measure or similarity function
Step 5:- Document clustering is done using k-means clustering algorithm

The email consists of structured information such as email header and unstructured information such as subject and body of the email. Text processing is done on unstructured information available in the message. Pre-processing of raw data is the first step for any email management task. In this step the email header, email body and attachments are parsed. From the parsed data, subject and email content or other fields are extracted. Once the required text is obtained, in our case it is email content or email body part, we need to remove stop words such as ‘the’, ’for’, ‘of’ etc. from the data obtained.

Next, stemming algorithm can be used to stem the data. For example, words ‘connection’, ’connecting’, ’connected’ will be converted to ‘connect’. After stopwords removal and stemming task, keywords or terms are identified. We can consider those nouns or pronouns which have higher frequency of occurrence. Once the data is extracted from the email, it is represented in some format. The most prevalent model for representation is the vector space model. In this model every email message is represented by a single vector and each element as a token or feature. In such type of data the tokens are usually words or phrases. These tokens can be categorized mainly into three categories- Unigram, Bigram and Co occurrence. Unigram are significant individual words. For example, in a data such as “Good Morning, my dear friends” the unigrams are ‘good’, ‘morning’, ‘my’, ‘dear’ and ‘friends’. Bigram are pair of two adjacent words. For example, in a data such as “hello friends, how are you?” the bigrams are ‘hello friends’, ‘friends how’, ‘how are’ and ‘are you’. In bigrams, word sequence is important, ‘hello friends’ and ‘friends hello’ are two different units. Co-occurrences are same as bigrams, only difference is, word sequence is not important. For example, ‘hello friends’ and ‘friend’s hello’ are treated as single unit as their order does not matter. There is another feature called target co occurrence which is same as co-occurrence with one target word inside each pair.

Once the term frequency is calculated, document vector is generated. For each email document, individual document vector is generated. Using document vector and similarity measure, similarity is calculated between email documents. The most similar documents are clustered together using clustering algorithm.

4. EXPERIMENTAL RESULTS

The effectiveness of email clustering using (SMTP) Similarity Measure for Text Processing [14], is tested by implementing k-means clustering algorithm with SMTP. The results are compared with k- means clustering algorithm implemented using other similarity measures- Euclidian distance, cosine similarity, extended jaccard coefficient and dice coefficient. We observed that, the results obtained by using SMTP with k-means clustering algorithm are better than other similarity measures. For implementation we used Intel(R) Core(TM) i3 processor 1.70 GHz with 4 GB RAM. The entire implementation is done using object oriented programming language Java.

4.1 Data set

For experimental purpose we used Enron email dataset which is the most popular email dataset and is available free on World Wide Web and can be downloaded from [16] .This dataset is cleaned and contains large set of email messages organized in folders and contains thousands of messages belonging to almost 150 users. For each user, one folder is allocated by the name of that user. Each of such folders contains subfolders such as inbox, sent, sent-items, drafts and other user created sub folders

.

The experimental results are based on emails contained inbox folder.

At initial stage, clusters are formed manually on the basis of similarity. For experimental purpose, four clusters are created and emails with similarity are put into four clusters.

Clusters generated by our system using k-means clustering algorithm are saved in a folder dynamically. Then this system generated clusters are compared with manually created clusters. The clusters returned by the system may not contain all the relevant mails; the clusters may contain some irrelevant mail. For example, out of 40 mails in a particular cluster returned by the system , only 30 mails would be relevant, means in real sense those 30 mails have similarity among themselves based on the mail content. Finally relevant mails in the system generated clusters are identified and accuracy is calculated as follows-

Suppose manually generated cluster contain emails {11,12,13,14}
System generated cluster contain emails{11,12,15}
This means matching contents are 2, so matchCounter is 2.
Accuracy=(matchCounter / (length of manually created cluster + length of system generated cluster - matchCounter)) *100
Substituting values in above formula, accuracy is calculated as –
(2 / (4 + 3 - 2)) *100=40%

4.2 Figures and Tables

Following table shows accuracy obtained for different clusters for five similarity measures. Overall results show that similarity measure SMTP retrieves maximum relevant emails.

Table 1. Accuracy obtained for different clusters for five similarity measures

Similarity Measure	**Cluster 1**	**Cluster 2**	**Cluster 3**	**Cluster 4**
Accuracy By Cosine Similarity	36.1111	6.4516	20.8333	28.5714
Accuracy By Euclidian Dist	24.6575	25.00	2.38095	20.00
Accuracy By Dice Coefficient	23.3333	8.8235	23.9130	3.8461
Accuracy By Extended Jaccard Coefficient	24.2424	20.00	27.4509	26.3157
Accuracy By SMTP	53.8461	23.0769	83.3333	31.5789

Following graphs show, the accuracy in percentage obtained for different clusters for five similarity measures.

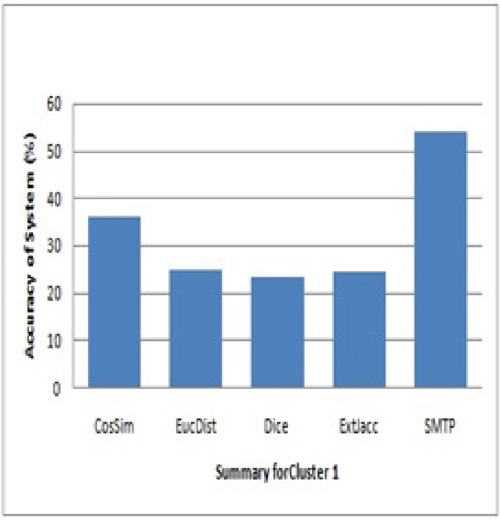


Figure 2. Accuracy for cluster 1

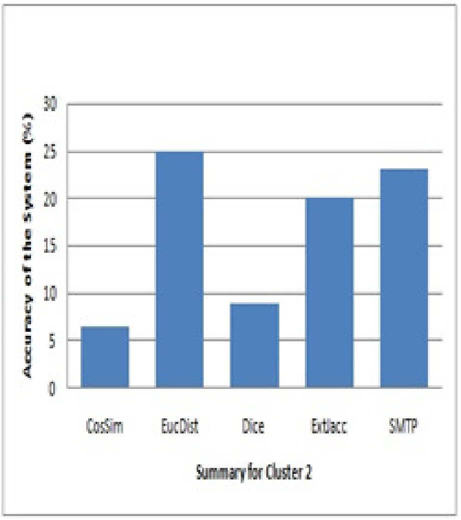


Figure 3. Accuracy for cluster 2

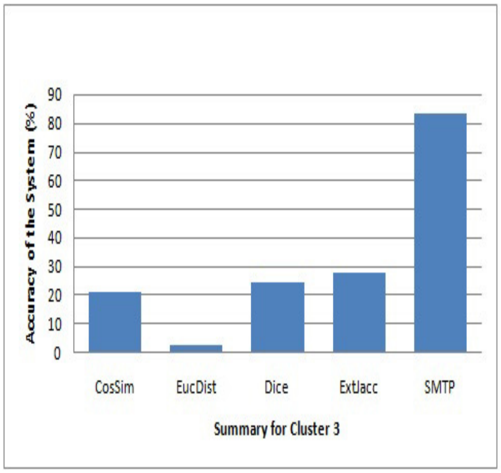


Figure 4. Accuracy for cluster 3

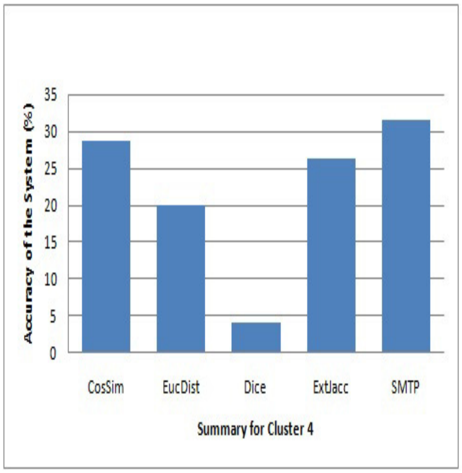
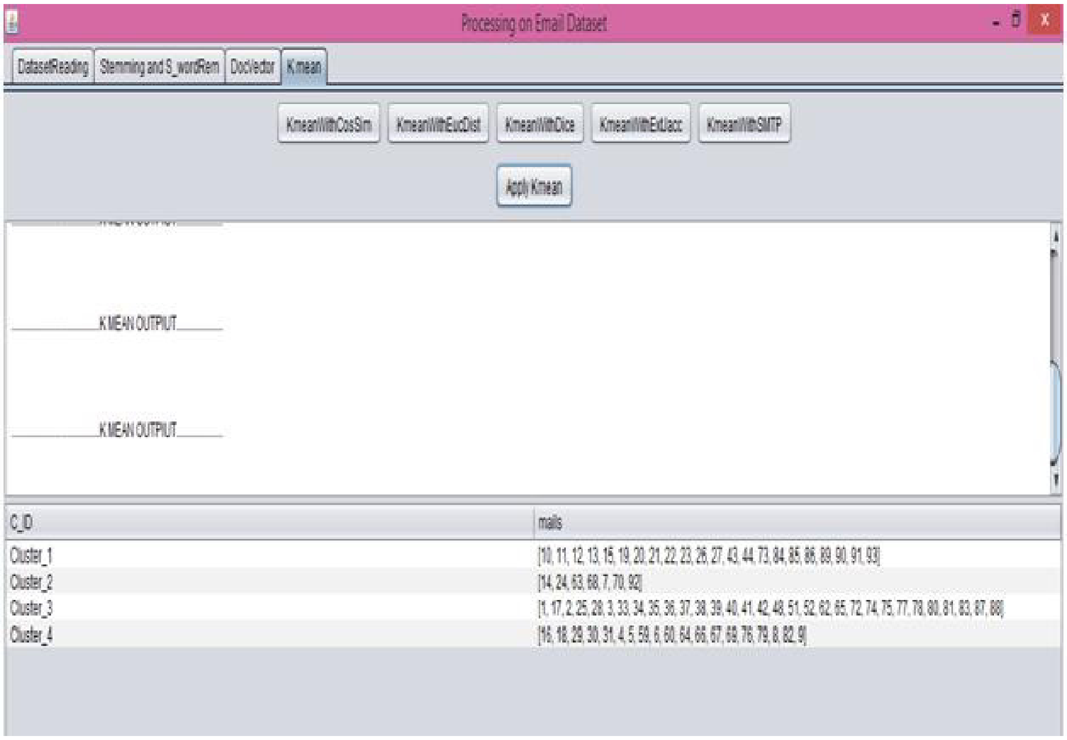


Figure 5. Accuracy for cluster 4

Following screen shot shows email clusters obtained using SMTP.



```
____________K MEAN OUTPUT____________

____________K MEAN OUTPUT____________
```

C_ID	mails
Cluster_1	[10, 11, 12, 13, 15, 19, 20, 21, 22, 23, 26, 27, 43, 44, 73, 84, 85, 86, 89, 90, 91, 93]
Cluster_2	[14, 24, 63, 68, 7, 70, 92]
Cluster_3	[1, 17, 2, 25, 28, 3, 33, 34, 35, 36, 37, 38, 39, 40, 41, 42, 48, 51, 52, 62, 65, 72, 74, 75, 77, 78, 80, 81, 83, 87, 88]
Cluster_4	[16, 18, 29, 30, 31, 4, 5, 59, 6, 60, 64, 66, 67, 69, 76, 79, 8, 82, 9]

Figure 6. Output for k- means Clustering Algorithm with SMTP

5. CONCLUSION

The effectiveness of email clustering is tested using k-means clustering algorithm with similarity measure SMTP. From the experimental results we have observed that SMTP gives better results for document (email) clustering than any other similarity measure used for the experimental purpose. This is mainly because SMTP considers presence or absence of features than difference between two values associated with present feature. So we can conclude that k-means clustering algorithm implemented with similarity measure SMTP gives best results for email (document) clustering.

On concluding note we suggest that, it would be very interesting to test how SMTP works with other clustering algorithms such as fuzzy c means clustering algorithm. Also it would be worth, analysing the SMTP on classification algorithms such as KNN.

ACKNOWLEDGEMENT

The authors wish to thank everyone who has contributed to this work directly or indirectly.

REFERENCES

[1] T. Joachims , (1997) "A Probabilistic analysis of the rocchio algorithm with TFIDF for text categorization", in *Proc. 14th Int.Conf. Mach. Learn.* ,San Francisco ,CA,USA, pp.143-151.

[2] H Kim, P. Howland & H. Park , (2005) "Dimension reduction in text classification with support vector machines", *J. Mach. Learn. Res.,* Vol. 6 pp. 37-53.

[3] G. Salton & M. J. McGill, (1983) *Introduction to Modern Retrieval.* London, U. K.: McGraw-Hill.

[4] J. Han & M. Kamber ,(2006) *Data Mining Concepts and Techniques,* 2nd ed. San Francisco ,CA, USA: Elsevier.

[5] T. W. Schoenharl & G. Madey , (2008) "Evaluation of measurement techniques for the validation of agent-based simulations against streaming data ",in *Proc. ICCS,* Krakow, Poland.

[6] C.G. Gonzalez , W. Bonventi, Jr. & A.L.V. Rodrigues, (2008) "Density of closed balls in real-valued and autometrized Boolean spaces for clustering applications", in *Proc. 19th Brazilizn Symp. Artif. Intel.,* Savador, Brazil, pp. 8-22.

[7] J. A. Aslam & M. Frost, (2003) "An information-theoretic measure for document similarity", in *Proc. 26th SIGIR,* Toronto, ON, Canada, pp. 449-450.

[8] D. Lin, (1998) "An information theoretic definition of similarity", in *Proc. 15th Int. Conf. Mach. Learn.,* San Francisco, CA,USA.

[9] J. D'hondt, J. Vertommen, P.A. Verhaegen, D. Cattrysse & R.J. Duflou, (2010) "Pairwise-adaptive dissimilarity measure for document clustering", *Inf. Sci.,* Vol. 180, No. 12, pp. 2341-2358.

[10] R.W. Hamming, (1950) "Error detecting and error correcting codes", *Bell Syst. Tech. J.,* Vol. 29, No. 2, pp. 147-160.

[11] S. Kullback & R.A.Leibler, (1951) "On information and sufficiency", *Annu. Math. Statist.,*Vol. 22, No. 1, pp. 79-86.

[12] G. H. Ball & D. J. Hall , (1967) "A clustering techniques for summarizing multivariate data", *Behav. Sci.,* Vol. 12, No. 2, pp. 153-155.

[13] R. O. Duda, P. E. Hart & D. J. Stork, (2001) *Pattern Recognition*, New York, NY, USA: Wiley

[14] Yung-Shen Lin, Jung-Yi Jiang & Shie-Jue Lee, (2014) "Similarity Measure for Text Classification and clustering", *IEEE Transactions on Knowledge and Data Engineering* , Vol. 26, No. 7.

[15] M. B. Eisen , P. T. S Spellman, P. O. Brown & D. Boststein, (1998) "A cluster analysis and display of genome-wide expression patterns", *Sci.,* Vol. 95, No. 25, pp. 14863-14868.

[16] https://www.cs.cmu.edu/~./enron/

Authors

Ms. Deepa B. Patil is a post graduate student in Computer Engineering from Savitribai Phule Pune University, Pune , Maharashtra State, India

Prof. Yashwant V. Dongre is Assistant Professor in Vishwakarma Institute of Information Technology (VIIT) at Savitribai Phule Pune University, Pune, Maharashtra State, India. His area of interest includes database management, data mining and information retrieval. He has several journal papers to his credit published in prestigious journals which includes AIRCCE IJDMS.