ENHANCING THE PERFORMANCE OF CLUSTER BASED TEXT SUMMARIZATION USING SUPPORT VECTOR MACHINE

M. S Patil[1], M. S. Bewoor[2], S. H. Patil[3]

[1]*Research Scholar, Department of Computer Engineering, BVUCOEP, Maharashtra, India*
[2]*Associate Professor, Department of Computer Engineering, BVUCOEP, Maharashtra, India*
[3]*Professor, Department of Computer Engineering, BVUCOEP, Maharashtra, India*

Abstract

Technology is evolving day by day and this increase in technology is nothing but is the efforts to reduce human work and to have systems as automatic as possible. Same thing is true in terms of existence of digital information. Due to enormous increase in the use of internet, there is striking increase in the digital information. This digital information is characterized by different form of information, same information in different form, unrelated information and also there is lot of redundant information. Another next important thing to note is that most of the time we require textual information. To search or retrieve small information one has to go through thousands of documents, read all the retrieved documents irrespective whether they contain useful information or no. It becomes very difficult to read all the retrieved documents and prepare exact summary out of it within time. Besides this, many times retrieved information is repeated in almost many documents. This leads to research in the area of text mining. Text summarization is one of the challenging tasks in the field of text mining.

Keywords: *Entropy, FCM, Purity, SVM*

----------***----------

1. INTRODUCTION

Text summarization is the process of presenting the information in the document in very precise manner without losing any information or content in the document. The approach that truly gives information contained in the document and in correct form without changing its meaning is considered to be the best approach. Thereof generated summary must retain the data as well as the central idea of document. Based on following characteristics, different text summarization techniques can be classified:

1. Based on number of documents (Single document and multi-document summarization.)
2. Based on summary generated (Extractive and Abstractive.)
   a. Extractive: Sentences in summary are same as those in the document.
   b. Abstractive: Sentences in summary are constructed from the information in the document. This approach is difficult as compared to extractive.
3. Based on technique used. (Supervised and unsupervised.)
4. Based on usage of summary (Query based and query independent.)
   a. Query based: Summary of the document is constructed with respect to the query given by the user.
   b. Query independent: This type of summary remains same throughout the process where sentences are selected from document irrespective of the query.

Irrespective of the type of summarization technique used, text summarization is carried out in following three stages:

1. Preprocessing
2. Processing
3. Summary Generation.

In the stage of preprocessing, NLP phases like tokenization, parsing, stop word removal, stemming, case folding etc are carried out. This stage eliminates unnecessary words and retains only important words.

In processing stage summarization algorithms are applied in order to extract sentences required for generating.

In last phase, final summary is generated from given document or documents.

This paper presents an extractive text summarization approach to generate summary. For this two algorithms are used viz. Fuzzy C Means (FCM), a clustering algorithm and Support Vector Machine (SVM). Finally summary generated is compared with the summary generated by the pure clustering algorithm.

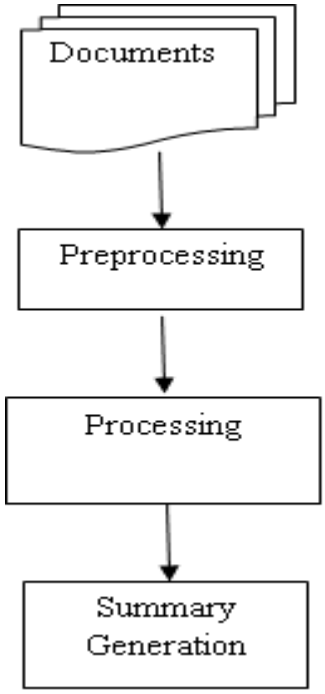

Fig-1: Summarization Process

2. RELATED WORK

The paper [2] proposes a system for generating summary using clustering algorithm cascaded with Support vector machine (SVM). It also proposes set of metrics for evaluating the performance of the proposed system with respect to performance of pure clustering algorithm. In Paper [3], author has compared various techniques of text summarization. In paper [7], Fuzzy C Means (FCM) clustering technique is described in detail. This paper clearly states the algorithm of FCM. In paper [6], performance of Fuzzy C Means (FCM) is compared with other techniques like Support Vector Machine (SVM), Artificial Neural Network (ANN) and BC. Also its results show, out of these three techniques SVM performs better. Use of SVM is also explained in this paper.

3. SYSTEM DESCRIPTION

Proposed system generates summary of text file using Fuzzy C Means (FCM), a clustering algorithm, cascaded with Support Vector Machine (SVM), a machine learning algorithm. First given input text file then undergoes preprocessing step which carries out NLP phases like tokenization, stop word removal, etc. Next is the processing step. In this FCM is applied and cluster centers are calculated. Then word count and word frequencies are calculated using FCM. Then in next step next algorithm is applied using SVM and word frequencies for SVM are calculated. In last step summaries are generated using two different sentence scores of the two algorithms viz. FCM and FCM cascaded with SVM. These summaries thus created are compared with respect to the set of metrics.

Following figure 2 shows architecture of the system.

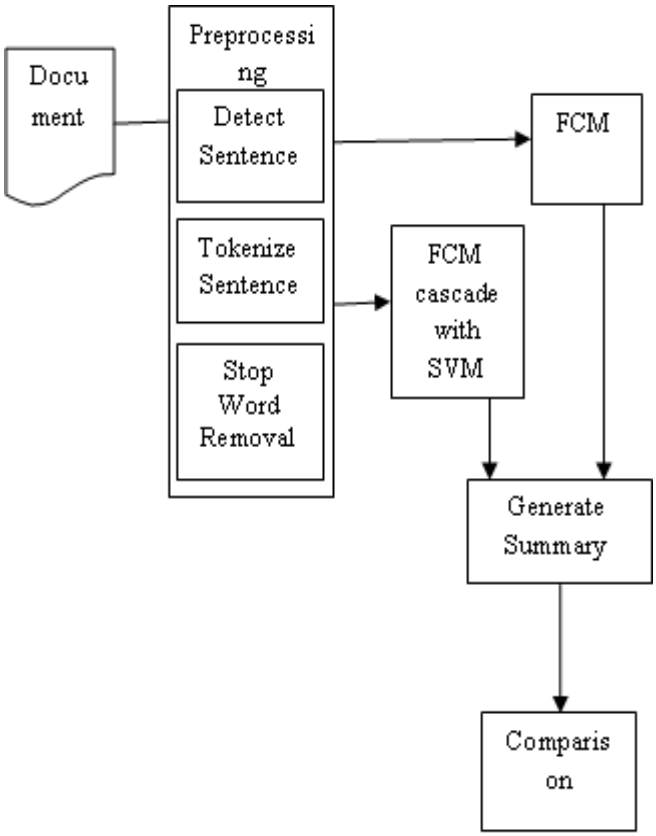

Fig-2: System Architecture

4. METHODOLOGY:

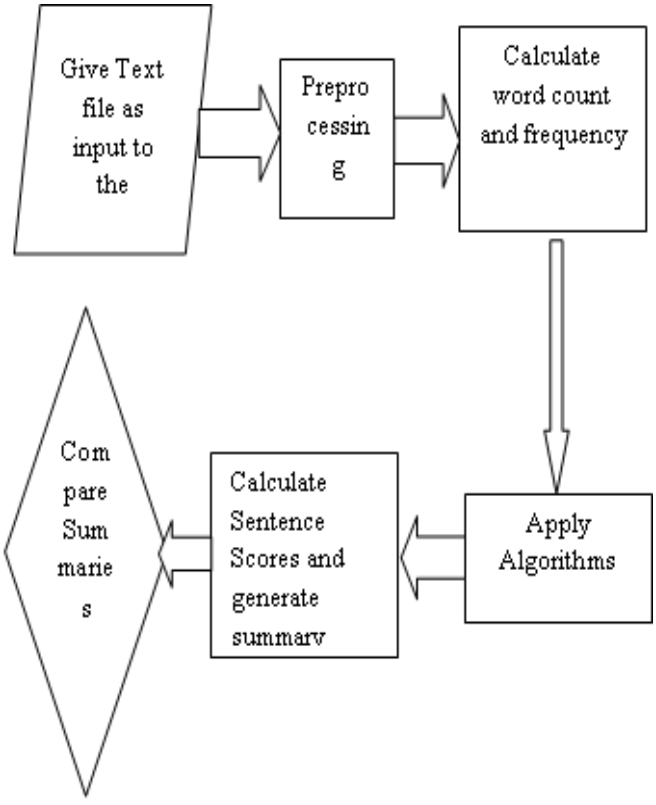

Fig-3: Workflow of the System

This paper proposes an algorithm for text summarization using FCM, clustering algorithm and SVM. Traditional clustering techniques like k-means, nearest neighbor clustering etc generate clusters in which each item belongs to exactly one cluster. These are termed as hard clustering techniques. Unlike these, FCM is a soft clustering technique. It allows one item to belong to all the generated clusters. Each item is related to all clusters with a relationship function. Higher the value of the function, higher is the relation of item with that cluster.

Work flow of the system is as follows:

1) Preprocessing will perform the NLP phases like tokenization, stop word removal, etc.
2) Next is to calculate word frequency as follows:

wf=((wordCount/TotalWords) *100)

where wordCount is the total number of times the word occur in the document.

Total Words is the total number of words in the document.

This frequency is normalized.

3) Then pure FCM algorithm is applied.

Minimizing function used for FCM is as follows:

$$\text{Uij}=1/\sum_{k-i}^{C} \frac{(dist(centeri,x)/}{(center2,x))}^{2/(m-1)}$$

Next to this SVM is applied. In this phase SVM kernel function is applied to calculate the sentence scores.

4) Sentence scores are calculated as follows:

Score= (X * Y) + C

Where
C is constant; here it is word frequency,
For FCM, X and Y are the cluster centers
For SVM, X and Y are calculated using the cluster centers and its related cluster values.

5) According to the limit of number of sentences in the summary, top high score sentences are selected and summary is generated.
6) Generated summaries are compared using following metrics:
   a. Purity: It is an external clustering evaluation metric. Higher the value of purity, more accurate the summary is obtained.

$$\text{Purity}= \frac{1}{N}(X+Y)$$

   b. Clustering Entropy: It is also an external clustering evaluation metric.

$$H(X) = -\sum_{i=0}^{n-1} \text{p(Xi)} * \text{log2 p(Xi)}$$

Here the value which is close to 0 is more accurate.

   c. Semantic Gap: This is shown by using sentence summary that shows semantic gap between two algorithms.
   d. Classification Cost: It is calculated using following formula:

Cost= Frequency + Overheads (Transaction or iterations) + number of clusters

5. EXPERIMENTAL AND PERFORMANCE ANALYSIS

Experiments have been performed by giving various text files as input to the system and calculating the values of above metrics using both the algorithms. Graphs have been generated for each metric showing the comparison between the values calculated using both the algorithms.

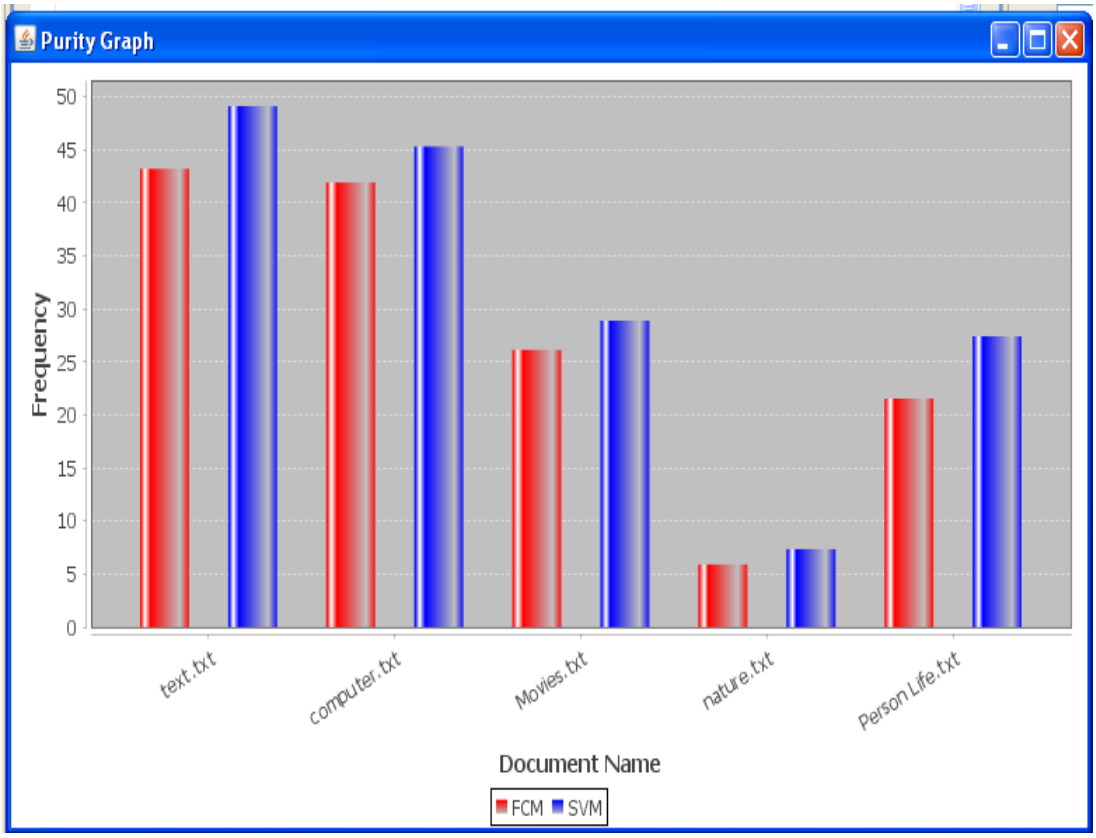

Fig-4: Comparison Graph of Purity in FCM and proposed algorithm

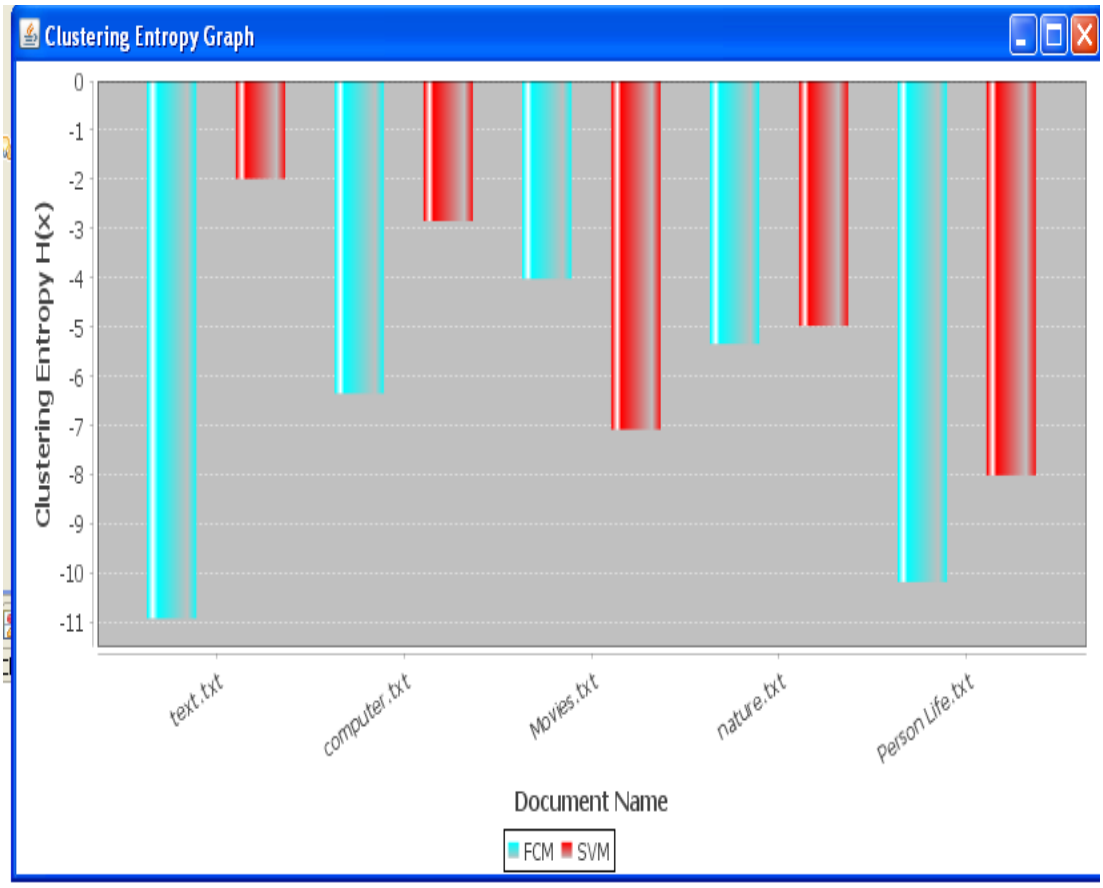

Fig 5: Comparison Graph of Clustering Entropy in FCM and proposed algorithm

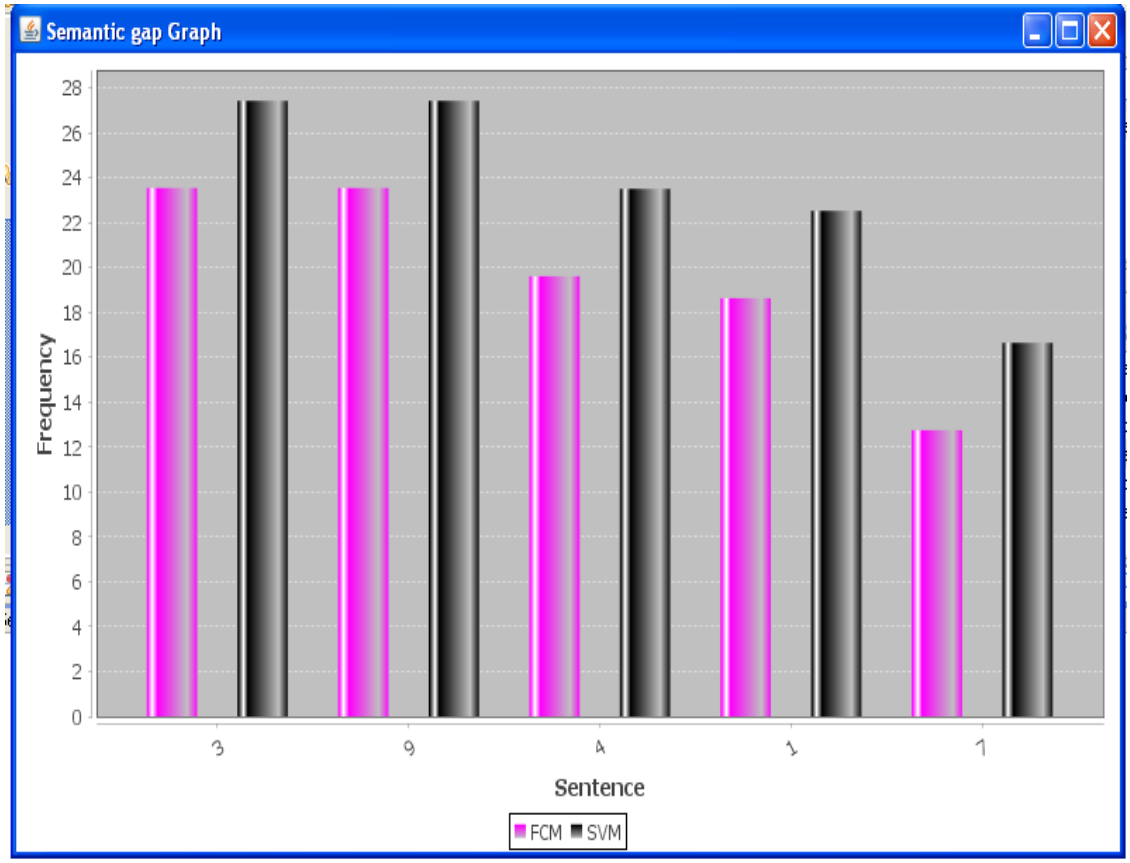

Fig 6: Comparison Graph of Semantic Gap in FCM and proposed algorithm.

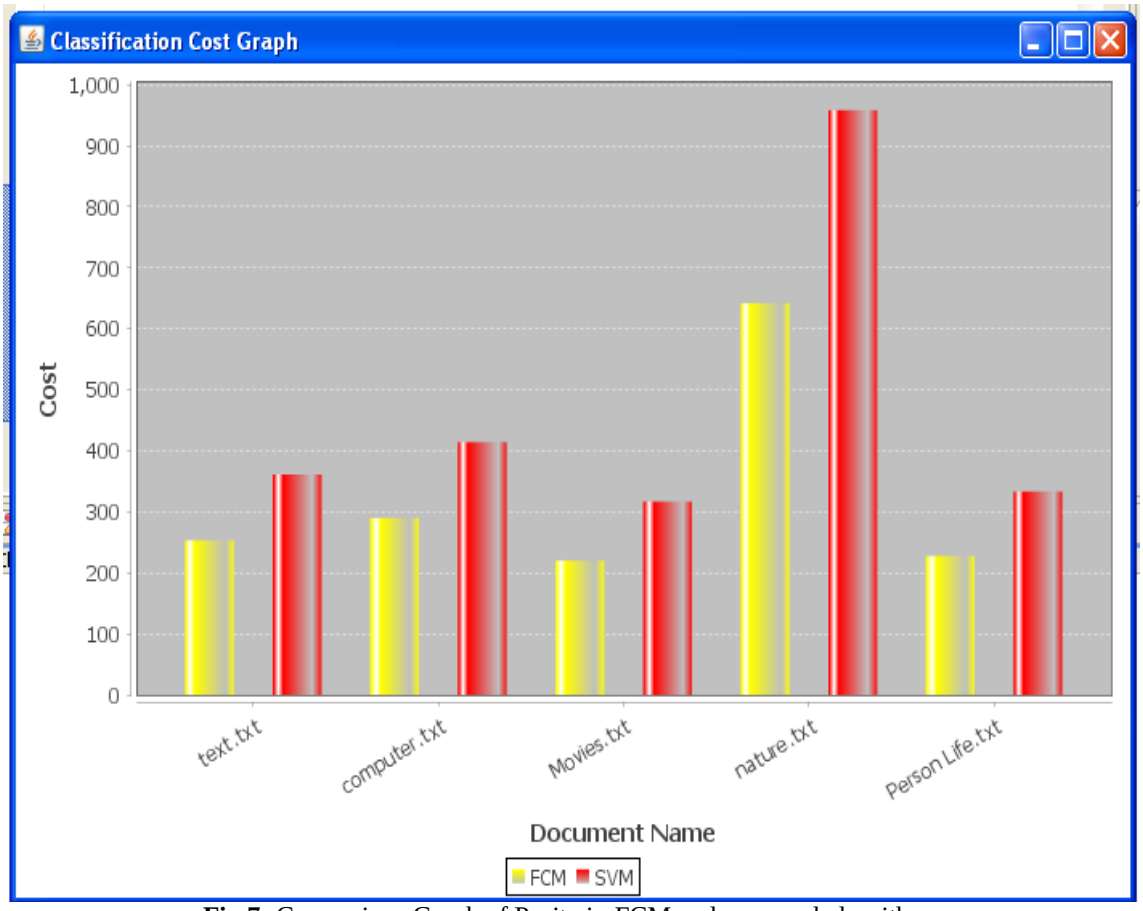

Fig 7: Comparison Graph of Purity in FCM and proposed algorithm.

6. CONCLUSION

In this paper concentration is given on improving the quality of summary generated by clustering technique. In this performance of the proposed system is compared with the performance of clustering technique. This is done by comparing the summaries generated on the basis of above given performance evaluating factors. Above results show that proposed algorithm performs better than pure clustering algorithm. Further this approach can be applied for multi-document summarization.

REFERENCES

[1]. A.K. Jain, M.N. Murty, P.J. Flynn, "Data Clustering: A Review", 2000 ACM
[2]. M. S. Patil, M. S Bewoor, Dr. S. H. Patil, "A Hybrid Approach for Extractive Document Summarization Using Machine Learning and Clustering Technique" IJCSIT 2014.
[3]. Roma V J, M S Bewoor, Dr. S. H. Patil, "the Quantity of NLP Based Text Summarization and Clustering Techniques By Quantitative and Qualitative Metrics" International Journal of Scientific & Engineering Research 2013
[4]. Ronen Feldman, James Sanger, "The Text Mining Handbook" www.cambridge.org
[5]. Ross, T. J. (2010); "Fuzzy Logic with Engineering Applications", Third Edition, John Wiley & Sons, Ltd, Chichester, UK
[6]. Srinivasa K G, Venugopal K R and L M Patnaik, "Feature Extraction using Fuzzy C - Means Clustering for Data Mining Systems" IJCSNS International Journal of Computer Science and Network Security, VOL.6 No.3A, March 2006
[7]. Sumit Goswami, Mayank Singh Shishodia, "A Fuzzy Based Approach To Text Mining And Document Clustering"
[8]. Tsutomu Hirao, Hideki Isozaki, Eisaku Maeda "Extracting Important Sentences with Support Vector Machines" ACM 2002.
[9]. Vishal Gupta, Gurpreet S. Lehal; "A Survey of Text Mining Techniques and Applications"; Journal of Emerging Technologies in Web Intelligence, Vol.1, No.1, August 2009.

BIOGRAPHIES

Patil M S is currently completing M.Tech(Computer) from Bharati Vidyapeeth College Of engineering Pune.
E-mail: madhsp.patil@gmail.com

M S Bewoor (M. E Computer) is currently working as an Associate Professor in the Department of Computer Engineering at Bharati Vidyapeeth College Of engineering Pune.
E-mail: msbewoor@bvucoep.edu.in

Dr. S H Patil is currently working as a Professor in the Department of Computer Engineering at Bharati Vidyapeeth College Of engineering Pune. His subject of specialization are Operating System and Distributed System.
E-mail: shpatil@bvucoep.edu.in