I J C T A, 9(40) 2016, pp. 267-274

A Wordnet Based Semantic Approach for Dimension Reduction in Multi label Text Documents

Rajni Jindal[*] and Shweta Taneja[**]

ABSTRACT

In this paper, we have proposed a new semantic approach for dimension reduction in multi label text documents. The proposed approach is based on the usage of Wordnet. There are three representations of documents we have considered: Bag of Words (BOW), Concise Semantic Analysis (CSA) and our proposed Semantic Analysis with Word net.

We have implemented our proposed approach on three multi label text datasets. These are collection of journal articles of computer science domain, Ohsumed dataset and Enron. The performance of the proposed approach is compared with the existing approaches using standard performance measures. These are: Recall, Precision, Micro average F1- score and Macro average F1- score. We have shown the results for two multi label classifiers: ML-KNN algorithm and BR algorithm. The results show that our proposed approach (Semantic Analysis with Wordnet) performs well as compared to the other existing approaches. On Computer Science journal articles, the best micro F1- score is 0.822, on Ohsumed corpus, the best micro F1- score is 0.700 and for Enron dataset, the best micro F1-score is 0.220 with our proposed approach.

Keywords: Text Categorization, Dimension Reduction, ML-KNN Algorithm, BR Algorithm

I. INTRODUCTION

The text categorization technique may be single label or multi label in nature. In single label, each text document may belong to one class, whereas in multi label, a text document may belong to more than one class. The real world text documents are multi label in nature. For example, classification of biological data, different types of reports etc. The text documents possess large size and huge number of features. This leads to a great challenge in categorization of text documents.

Dimension reduction [1, 2] is a method used to reduce the features as well as the dimensions of text documents. It is used in many applications like image recognition problems, biological data etc. Many researchers have contributed in this field. Feature selection [3] and feature reduction [4, 5] are concepts used in dimension reduction. Feature selection selects a subset of features; Feature reduction reduces the dimensionality by combining certain features. The choice of the method depends on the application domain and the type of problem.

In this paper, we have proposed a new semantic approach for dimension reduction in text documents. Our approach is based on the use of Wordnet. The concept of hypernyms of words is used in the proposed approach. Firstly, we tokenize the text documents using a tokenizer. Then the hypernyms of generated tokens are drawn. The similarity between the tokens is found by detecting the common hypernym. In this way, semantic relatedness between the tokens is detected. As a result, we get a reduced number of tokens.

[*] Computer Engineering Department, Delhi Technological University, New Delhi 110042, India, *E-mail: rajnijindal@dce.ac.in*

[**] Research Scholar, Computer Engineering Department, Delhi Technological University, New Delhi - 110042, India, *E-mail: shweta_taneja08@yahoo.co.in*

The organization of the paper is as follows: Section 2 presents the related work done in dimension reduction. In Section 3, the framework of the proposed approach is given. Section 4 explains the working of the proposed approach with the help of an example. The details of experiments conducted, results obtained and performance comparison of the proposed approach is discussed in section 5. The next section concludes the work.

II. RELATED WORK

Different researchers have contributed and suggested their methods for dimension reduction. In our proposed algorithm, we have used semantics of the tokens (using Word Net) to reduce the number of tokens. Table1 shows a summary of work done by different authors in dimension reduction and characteristics of our approach.

Table 1
Summary of work done in dimension reduction by different authors and features of our proposed approach

Authors	*Problem Addressed*	*Proposal*	*Dataset*	*Results*
S. Deerwester (1990) [7]	Dimension Reduction	Singular value decomposition (SVD), based on decomposing a large term by document matrix	MED-medical abstracts, CISI-information science abstracts	Results show it is a promising method
A. S. Ramkumar, Dr. B. Poorna (2016) [8]	Dimension Reduction	Proposed Document Clustering Using Dimension Reduction	BBC Sports Dataset	This method shows significant improvement in Accuracy, Precision and Recall.
Hyunsoo Kim *et al.* (2005) [9]	Dimension Reduction	Support Vector Machines are used for reducing the dimensions of documents	MEDLINE dataset, Reuters-21578	It achieves better efficiency in both training and testing the data.
Evgeniy Gabrilovich, Shaul Markovitch (2009)[10]	Semantic Interpretation of Natural Language texts	Explicit Semantic Analysis technique (ESA)	Used concepts derived from Wikipedia	Significant improvements over existing Algorithms
Chenping Hou *et al.* (2010) [11]	Dimension Reduction	Constraints are used for multiple view dimension reduction problems.	WebKB, 20-News-Group and Sonar data	Their approach outperforms other approaches.
Li Zhixing *et al.* (2011) [12]	Dimension Reduction	Concise semantic Analysis technique	Reuters-21578, 20-News-Group and Tancorp	Their approach reaches a comparable performance with SVM.
Koushik Mallick and Siddhartha Bhattacharyya (2012) [13]	Dimension Reduction	Distance between data points is counted and scatter matrix is calculated.	Reuters dataset	Their approach is more efficient than other state of art algorithms.
Hu Guan *et al.* (2013) [14]	Dimension Reduction	Imprecise Spectrum Analysis for fast dimension reduction	WebKB, Reuters-21578 and 20-News-Group	Their approach achieves fast and competitive classification accuracy with state of art algorithms.
Our Proposed Approach	Dimension Reduction	Semantic Analysis using Word Net	Ohsumed dataset, Computer Science journal dataset and Enron dataset	Our approach achieves a significant reduction in the number of tokens.

III. FRAMEWORK OF PROPOSED APPROACH

The following figure 1 shows the framework of the proposed approach.

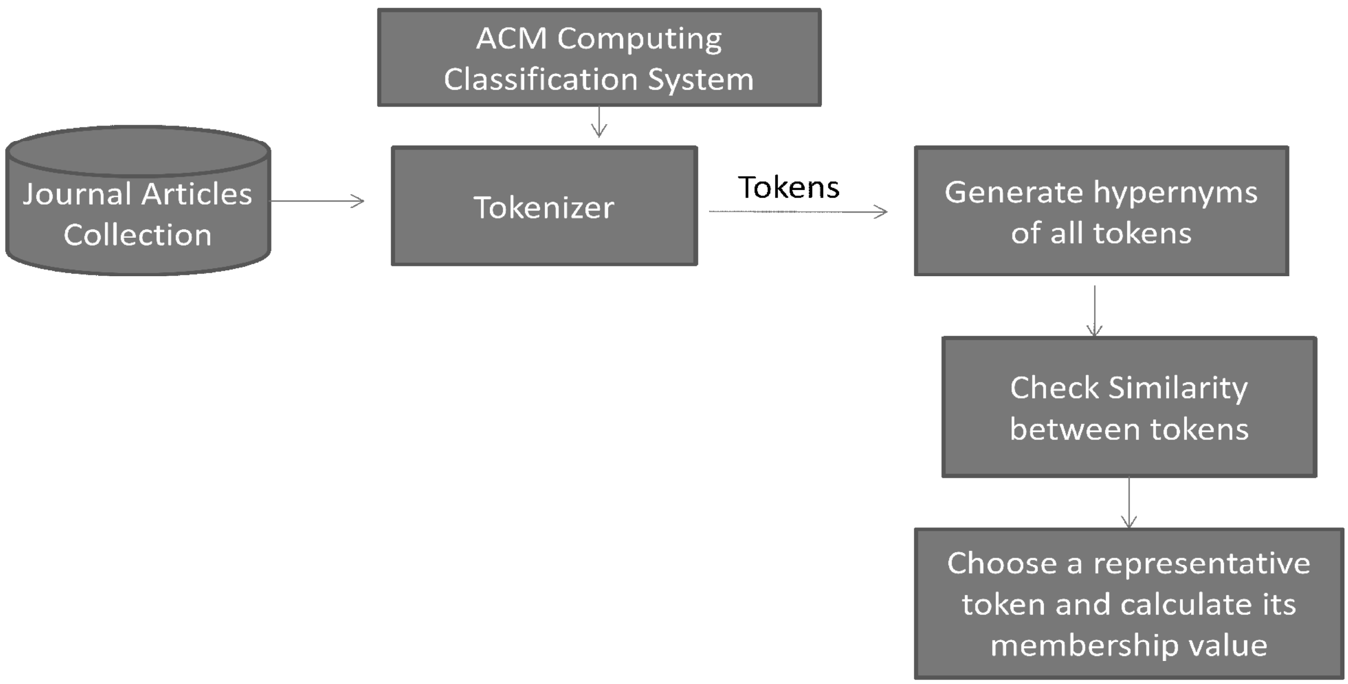

Figure 1: Framework of proposed approach

A collection of journal articles is taken as input. The first module TOKENIZER scans the Abstract, Title and Keywords of the input journal article. It identifies tokens using the standard ACM Computing Classification System, 2012[6]. Using Wordnet, the hyperyms of the tokens are drawn upto four levels. The tokens having common hypernym are identified. A representative is chosen between the two tokens. It is that token which has more frequency. The frequency of the selected token is increased by the average of the two frequencies. This process is repeated. Finally we get a reduced number of tokens.

IV. AN EXAMPLE SHOWING THE PROPOSED APPROACH

To show the working of proposed approach, let us take an example. Suppose we take a **journal article (**belonging to Computer Science domain**) titled** “Mining Network data for Intrusion Detection through combining SVMs with ant colony Networks”. The **keywords** are: Data mining, Data classification, Intrusion detection system (IDS), Machine learning, Support vector machine, Ant colony optimization. The **abstract** of the article is given in figure 2.

In this paper, we introduce a new machine-learning-based data classification algorithm that is applied to network intrusion detection. The basic task is to classify network activities (in the network log as connection records) as normal or abnormal while minimizing misclassification. Although different classification models have been developed for network intrusion detection, each of them has its strengths and weaknesses, including the most commonly applied Support Vector Machine (SVM) method and the Clustering based on Self-Organized Ant Colony Network (CSOACN). Our new approach combines the SVM method with CSOACNs to take the advantages of both while avoiding their weaknesses. Our algorithm is implemented and evaluated using a standard benchmark KDD99 data set. Experiments show that CSVAC (Combining Support Vectors with Ant Colony) outperforms SVM alone or CSOACN alone in terms of both classification rate and run-time efficiency.

In this paper, we introduce a new machine-learning-based data classification algorithm that is applied to network intrusion detection. The basic task is to classify network activities (in the network log as connection records) as normal or abnormal while minimizing misclassification. Although different classification models have been developed for network intrusion detection, each of them has its strengths and weaknesses, including the most commonly applied Support Vector Machine (SVM) method and the Clustering based on Self-Organized Ant Colony Network (CSOACN). Our new approach combines the SVM method with CSOACNs to take the advantages of both while avoiding their weaknesses. Our algorithm is implemented and evaluated using a standard benchmark KDD99 data set. Experiments show that CSVAC (Combining Support Vectors with Ant Colony) outperforms SVM alone or CSOACN alone in terms of both classification rate and run-time efficiency.

Figure 2: Abstract of Sample Article

The following table 2 shows a list of tokens generated by the Tokenizer module. There is a list of 14 tokens along with their frequency.

Table 2
Tokens with their frequency of the Sample Article

S.no	*Tokens*	*Frequency*
1	Machine	6
2	Learning	3
3	Data	8
4	Classification	6
5	Algorithm	2
6	Network	7
7	Intrusion	6
8	Detection	5
9	Ant	4
10	Mining	4
11	Colony	4
12	Optimization	4
13	Support	5
14	Vector	5

Using Word net, the hypernyms of the tokens are drawn upto four levels. These are called as hypernym trees of the tokens. It is shown in following table3.

Table 3
Hypernym trees of the tokens

Tokens	*L1*	*L2*	*L3*	*L4*
Machine	*Device*	*Internet*	*Artifact*	*Whole*
Learning	*Cognitive process*	***Process***	*Psychological factor*	-
Data	*Information*	*Cognition*	*Psychological factor*	-
Classification	*Grouping*	*Activity*	*Act, human action*	*Event*
Algorithm	*Rule*	***Process***	*Procedure*	-
Network	*System*	*Group*	-	-
Intrusion	*Entrance*	*Arrival*	*Action*	-
Detection	*Perception*	*Cognitive process*	*Process*	*cognition*

Tokens	*L1*	*L2*	*L3*	*L4*
Ant	*Insect*	*Insect*	*Arthropod*	-
Mining	*Production*	*Industry*	*Business*	*Commerce*
Colony	*Animal group*	*Biological group*	*Group*	-
Optimization	*Improvement*	*Change*	*Art*	-
Support	*Activity*	*Human Action*	*Event*	-
Vector	*Variable*	*Quantity*	*Concept*	-

As shown in above table 3, the tokens LEARNING and ALGORITHM have the common hypernym PROCESS at same level. So these two tokens are similar. The token learning has more frequency than the token ALGORITHM. So, token LEARNING is selected and its frequency becomes as 6 (3+ average (3, 2)). This process is repeated for all the tokens. Finally we get a reduced set of tokens.

V. EXPERIMENTS CONDUCTED AND RESULTS OBTAINED

5.1. Datasets Used

In our experiments, we have taken three multi label text datasets. First is a collection of 200 journal articles of computer science domain. These journal articles are selected in such a way that they belong to different sub domains under Computer Science like: data mining, computer networks, cryptography etc. Second dataset is a subset of MEDLINE database maintained by National Library of Medicine. It is called as Ohsumed dataset [15]. We have used the dataset used by Joachims in 1998[1]. The third dataset is Enron [16]. It contains email messages. It is a subset of UC Berkeley Enron Email Analysis Project. The detail of datasets is given in following table 4.

Table 3
Details of Dataset

Name	*Instances*	*Labels*	*Attributes*	*Cardinality*	*Density*
Computer Science Journals Dataset	200	10	3	2.5	0.17
Ohsumed test corpus	13929	23	1025	1.66	0.07
Enron	1702	53	1001	3.378	0.064

5.2. Results and Performance Comparison of Proposed Approach

To compare the performance of the proposed approach with the existing approaches, standard performance measures are used. These are: Recall, Precision, Micro average F1- score and Macro average F1- score. Precision is the rate of how many machine-labeled positive samples are truly positive samples and recall is the rate of how many truly positive samples are given a positive label by a machine learning algorithm. F1- score is a geometric mean of both recall and precision. If there is a single category, F1- score is sufficient. But for the whole dataset, Micro- average F1- score and Macro-average F1- score is needed. The macro-average F1 is the average of all F-1 values and micro-average F1 is the combination of global precision and recall.

The most common method of representing text documents is Bag of Words (BOW). In BOW representation [18], features are words and text documents are considered as a collection of words. The frequency of words is taken into account. This representation has drawbacks. The text documents possess high dimensionality. This representation faces a great challenge due to this problem. Second is the semantics of words. This representation does not consider the similarity and other relationships between words. It simply treats a document as a collection of words. We have compared BOW approach with our proposed approach. There are three representations: BOW, Concise Semantic Analysis (CSA) [19] and our proposed Semantic Analysis with Word net. We have shown the results for two multi label classifiers: ML-KNN algorithm and BR algorithm.

The proposed approach is implemented using a WEKA-based framework (WEKA tool) running under Java JDK 1.6 along with the libraries of MEKA [17] and Mulan. Experiments are conducted on 64 bit machines with 2.6 GHz of clock speed. Evaluation is done in the form of training and test split on each dataset. The split into training and test is done on a random basis and repeated multiple times. 10 fold cross validation is used each time. The performance comparison of the proposed approach with the other approaches is shown in the table 5 given below.

Table 5
Performance Comparison

DATASET	*APPROACH*	*ML-KNN*		*BR*	
		Micro F1-Score	*Macro F1-Score*	*Micro F1-Score*	*Macro F1-Score*
Computer Science Articles	**BOW**	0.766	0.710	0.734	0.700
	CSA.	0.772	0.743	0.745	0.710
	S.A.+ Wordnet	**0.822**	**0.791**	**0.762**	**0.711**
Ohsumed dataset	**BOW**	0.623	0.591	0.585	0.550
	CSA	0.655	0.622	0.611	0.573
	S.A.+ Wordnet	**0.700**	**0.654**	**0.634**	**0.592**
Enron	**BOW**	0.172	0.100	0.155	0.050
	CSA	0.182	0.132	0.176	0.122
	S.A.+ Wordnet	**0.220**	**0.192**	**0.201**	**0.166**

From the above table, it is clear that our proposed approach (S.A with Wordnet) performs well as compared to the other existing approaches. On Computer Science journal articles, the best micro F1- score is 0.822 where semantic analysis with Word net is used. On Ohsumed corpus, the best micro F1- score is 0.700 and for Enron dataset, the best micro F1- score is 0.220. Also, the results are shown graphically in following figures 3,4 and 5 for all the three datasets.

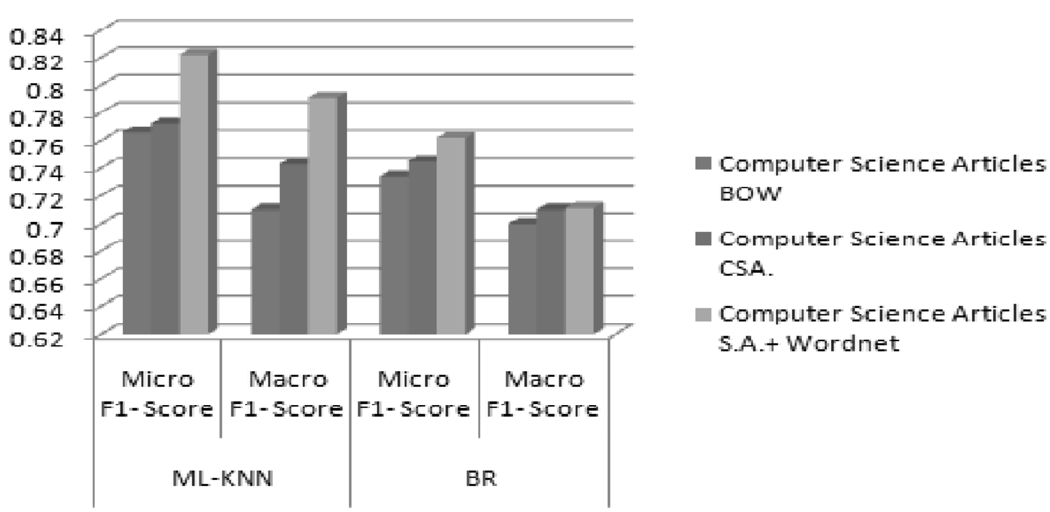

Figure 3: Performance Comparison on Computer Science Articles dataset

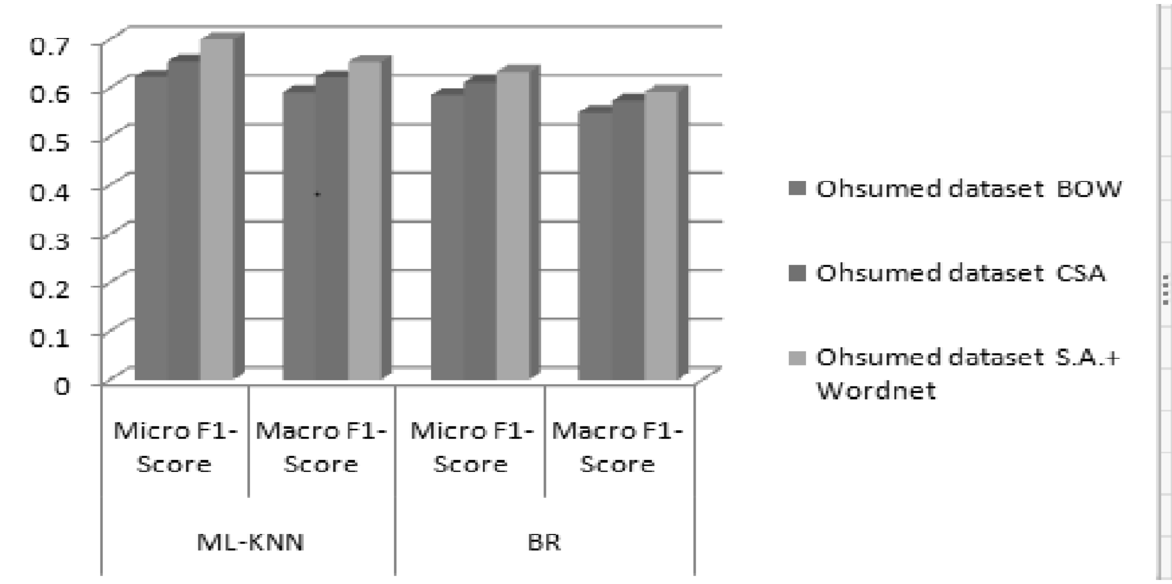

Figure 4: Performance Comparison on Ohsumed dataset

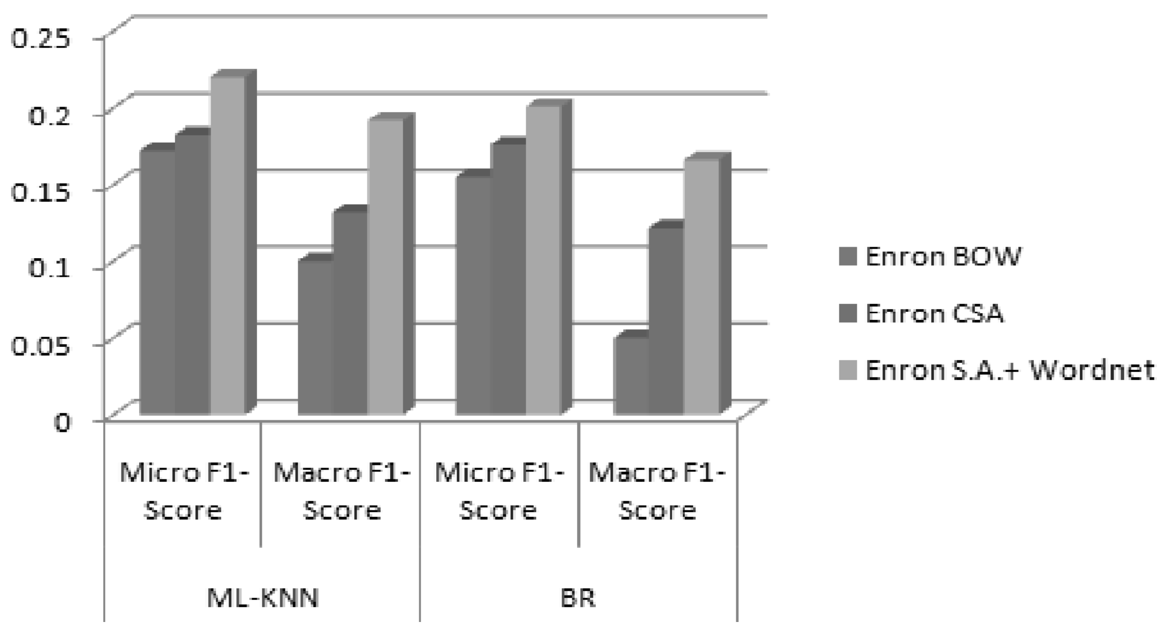

Figure 5: Performance Comparison on Enron dataset

VI. CONCLUSION

In this paper we have proposed a semantic approach using Wordnet for dimension reduction in multi label text documents. We have implemented the proposed approach on three multi label text datasets. The performance is evaluated using standard measures like Recall, Precision, Micro average F1- score and Macro average F1- score. There are three representations of documents we have considered: BOW, Concise Semantic Analysis and our proposed Semantic Analysis with Word net. We have shown the results for two multi label classifiers: ML-KNN algorithm and BR algorithm. The results show that our proposed approach (Semantic Analysis with Wordnet) performs well as compared to the other existing approaches. On Computer Science journal articles, the best micro F1- score is 0.822, on Ohsumed corpus, the best micro F1- score is 0.700 and for Enron dataset, the best micro F1- score is 0.220.

REFERENCES

[1] T. Joachims , “Text categorization with support vector machines: Learning with many relevant features”, Tenth European conference on machine learning, pp. 137-142, 1998.

[2] S.Godbole and S. Sarawagi, “Discriminative methods for multi-labeled classification”, 8th Pacific- Asia Conference on Knowledge Discovery and Data Mining, 2004.

[3] D. Koller, M. Sahami, “Toward optimal feature selection, in: International Conference on Machine Learning” (ICML), pp. 284–292, 1996.

[4] S. Mika, G. Ratsch, J. Weston, B. Scholkopf, K.R. Mullers, Fisher discriminant analysis with kernels, in: Neural Networks for Signal Processing IX, 1999. Signal Processing Society Workshop, pp. 41–48, 1999.

[5] K. Pearson, On lines and planes of closest fit to systems of points in space, Philosophical Magazine, vol. 2, issue 6, pp. 559–572, 1901.

[6] ACM Computing Classification System, [online classification].Available at: http://delivery.acm.org/10.1145/2380000/2371137/ACMCCSTaxonomy.html?, 2012.

[7] S. Deerwester et al., “Indexing by latent semantic analysis,” Journal of the Society for Information Science, vol. 41, issue 6, pp. 391-407, 1990.

[8] A. Sudha Ramkumar, Dr. B. Poorna, “Text Document Clustering Using Dimension Reduction Technique”, International Journal of Applied Engineering Research, vol. 11, issue 7, pp 4770-4774, 2016.

[9] Hyunsoo Kim et. al.,” Dimension Reduction in Text Classification with Support Vector Machines”, Journal of Machine Learning Research, vol. 6, pp. 37-53, 2005.

[10] E. Gabrilovich, S. Markovitch , “Wikipedia-based Semantic Interpretation for Natural Language Processing”, Journal of Artificial Intelligence Research, vol. 34, pp. 443-498, 2009.

[11] C. Hou et. al.,” Multiple view semi-supervised dimensionality reduction”, Journal of Pattern Recognition Letters, Elsevier, vol. 43, pp. 720-730, 2010.

[12] L. Zhixing et al., “Fast text categorization using concise semantic analysis”, Journal of Pattern Recognition Letters, vol. 32, pp. 441-448, 2011.

[13] K. Mallick, Siddhartha Bhattacharyya,” Uncorrelated Local Maximum Margin Criterion: An Efficient Dimensionality reduction Method for Text Classification”, Journal of Procedia Technology, vol. 4, pp. 370-374, 2012.

[14] H. Guan et. al., “Fast dimension reduction for document classification based on imprecise spectrum analysis”, Journal of Information Sciences, vol. 222, pp. 147-162, 2013.

[15] Hersh W. et al., ’Ohsumed: An interactive retrieval evaluation and new large text collection for research’, 17th ACM International Conference Research and Development in Information Retrieval, pp. 192-201, 1994.

[16] Enron dataset. [Online] http://mulan.sourceforge.net/datasets.html.

[17] Meka tool [Online] http://MEKA.sourceforge.net

[18] M. Lan, M. et al.,”Supervised and traditional term weighting methods for automatic text categorization. IEEE Trans. Pattern Anal. Machine Intell. , vol. 31, issue 4, pp. 721–735, 2009.

[19] Li Zhixing et al.,” Fast text categorization using concise semantic analysis”,Pattern Recognition Letters, vol. 32, pp. 441-448, 2011.