Volume 3, No. 7, Nov-Dec 2012
ISSN No. 0976-5697
International Journal of Advanced Research in Computer Science
REVIEW ARTICLE
Available Online at www.ijarcs.info

Conceptually Co-occurring Words Included as Feature Selection in Text Document Classification using SVD and SVM

D. Malathi*
Department of CA
Bannari Amman Institute of Technology
Sathyamangalam, Tamilnadu, India
malathisubbu@gmail.com

Dr. S. Valarmathy
Department of ECE
Bannari Amman Institute of Technology
Sathyamangalam, Tamilnadu, India
atrmathy@gmail.com

Abstract: Document classification is a means of knowledge extraction in text mining process. This has been experimented by so many researches. But still we have included one of the extra features in the preprocessing phase and checked its outcome with Support Vector Machine (SVM). The feature selection has been accounted with the Vector space method, Single Value Decomposition (SVD) which is specifically meant for dimension reduction. The experiment has shown some positive changes in classification.

Keywords: Document classification – Support Vector Machine – Singular Value Decomposition

I. INTRODUCTION

Text mining works with unstructured or semi-structured data sets. This ranges from emails, web pages, newspaper articles, market research reports, curriculum vitae, complaint letters from customers, intranet communication, workgroup communication, social media analysis, short message services, legal documents, anonymous notes, internally generated reports and so on. The technologies that have been developed and can be used in the text mining process are search engines, information extraction, topic tracking, summarization, categorization, clustering, concept linkage, information visualization, sentiment analysis and question answering [1]. Text categorization is the problem of automatically assigning predefined categories to free text documents [2].

The document classification, which handles the string of characters, has to be represented that is best suitable for the learning and the classification algorithm. Information Retrieval concentrates on words as a representation units and their ordering in a document is of less importance for many tasks, strategy leads to the representation of documents as bags of words [3].

Document classification is done by sequential steps of sentence splitting, tokenization, part-of-speech, tagging, stop word list, stemming, noisy data, word sense, collocations, syntax, text representation, Indexing, upper ontology, domain ontology. Out of these, steps may be altered or skipped according to the need of outcome [4].

The feature space of document data consists of unique terms such as words and phrases. A moderate sized text collection consists of tens and hundreds of terms. The major drawback in the text categorization problems is the high dimensionality [2] of the feature space.

The process of pattern finding involves observation in varied dimensions. This is due to increase in variables associated with each observation. High-dimensional datasets present many mathematical challenges and are bound to give rise to new theoretical developments [5]. One of the problems with high-dimensional datasets is that, in many cases, not all the measured variables are important for understanding the underlying phenomena of interest. While certain computationally expensive novel methods [6] can construct predictive models with high accuracy from high-dimensional data, it is still of interest in many applications to reduce the dimension of the original data prior to any modeling of the data.

Statistical composition of patterns from text can be learned from Dimensionality Reduction methodologies. Dimensionality reduction is the transformation of high-dimensional data into a meaningful representation of reduced dimensionality manifold.

A. *Dimensionality Reduction:*

The problem of (nonlinear) dimensionality reduction can be defined as follows. Assume a dataset represented in a n × D matrix X consisting of n data vectors xi ($i \in \{1, 2, \ldots, n\}$) with dimensionality D. Dimensionality reduction techniques transform dataset X with dimensionality D into a new dataset Y with dimensionality d, while retaining the geometry of the data as much as possible [8].

i.e., Input: $X = (x_1, \ldots, xn) : xi \in R^D$

Output: $Y : y_j \in R^d$ *such that* $d << D$

II. LITERATURE SURVEY

Literature Survey on Dimension Reduction Methodologies has been done in detail by the authors in [7]. Document classification methodologies have been studied in vide course to procure knowledge on evaluation schema of data reduction.

The observed papers in [7] have motivated to proceed document classification as mentioned in Figure 1. The procedure identified for document classification is mentioned below:

a. Documents from various sources are collected and processed to remove stop words and stemming.
b. Term – Frequency matrix is obtained and normalized
c. Dimension Reduction technique is applied to the above mentioned matrix identify relevant feature for document classification

d. Machine learning algorithm is applied on feature extracted to classify the documents
e. The classification can be used on varied applications of IR.

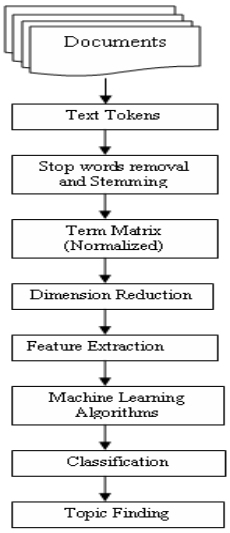

Figure 1: Document Classification Procedure

A text document is defined as, $D = \{d_1, d_2, \ldots, d_n\}$ be a document set of n documents, where $d_1, d_2, \ldots, d_n$ are individual documents and each document belongs to one of the classes in the set $\{c_1, c_2, \ldots, c_p\}$[9]. Each document can be categorized to two or more classes. Let the word set $W=\{w_1, w_2, \ldots, w_m\}$ be the feature vector of the document set. Then each document d_i, $1 \le i \le n$, is represented as $d_i = \langle d_{i1}, d_{i2}, \ldots, d_{im} \rangle$, where each d_{ij} denotes the number of occurrences of word w_j in the i^{th} document. Purposely sequence co-occurrence of words is included for features selection of concept by word count method. For example stop word does not include pronouns. This might give impact in rise in frequency of conceptual words.

A. Dataset:

The experiment is carried out using Reuters-21578 dataset [10]. Document distribution over categories in both the training and the testing sets is balanced, i.e., for experimental set up training phase has been included with 200 documents of 5 categories and testing phase has been included with 500 documents of same 5 categories and 1 new category.

B. Text Document Preprocessing:

To obtain data for classification documents are cluttered into collection of conceptual terminologies of words that are used in a by a *tokenization* process, i.e. a text document is split into a stream of words by removing all punctuation marks [11] and by replacing tabs and other non-text characters by single white spaces and stemming is done using the standard Porter's algorithm [12]. This tokenized representation is then used for further processing. In order to reduce the size of the set of words describing document can be reduced by filtering (stop word removal) and stemming.

C. Vector Space Model (VSM):

The VSM represents documents as vectors in *m* dimensional space, i.e. each document *d* is described by a numerical feature vector $w(d) = (x(d, t1), \ldots, x(d, tm))$. The main task of the vector space representation of documents is to find an appropriate encoding of the feature vector. Each element of the vector usually represents a word (or a group of words) of the document collection, i.e. the size of the vector is defined by the number of words (or groups of words) of the complete document collection. The simplest way of document encoding is to use binary term vectors, i.e. a vector element is set to one if the corresponding word is used in the document and to zero if the word is not.

Vector space model is useful because it provides an efficient, quantitative representation of each document. VSM contains the number of occurrences of word *j* in document *I*, say, f_{ji}, and the number of documents which contain the word *j*, say, d_j. A common approach uses the solution, f_{ji} X d_j matrix. Using these counts, [13] we can represent the i^{th} document as a *w*-dimensional vector x_i as follows. For $1 \le j \le w$, set the j^{th} component x_i, to be the product of three terms

$$x_{ij} = t_{ji}\, g_i\, s_i$$

Where t_{ji} is the *term weighing component* and depends only on f_{ji} , while g_j is the *global weighing component* and depends on d_j and s_i is the *normalization component* for x_i. t_{ji} captures the relative importance of a word in a document, while g_j captures the overall importance of a word in the entire set of documents. The objective of such weighing schemes is to enhance discrimination between various document vectors for better retrieval effectiveness. In this paper we use the *term frequency-inverse document frequency*. The scheme is

$$t_{ji} = f_{ji},\ g_j = \log(d/d_j) \text{ and } s_i = 1/(t_{ji}\ (t_{ji}\ g_j)2)^{-½}$$

III. CLASSIFIER MODEL

A. Singular Value Decomposition:

Singular Value Decomposition [14] is a mathematical method, which says that a rectangular matrix A= f_{ji} X d_j be a document-term matrix with positive real value entries. The rank reduced singular value decomposition is performed on the matrix to determine patterns in the relationship between the term and concepts contained in text. This decomposition is defined as

$$A = U \Sigma V^T$$

$j \times r$ $\quad r \times r$ $\quad V$ is $r \times i$

The columns $\boldsymbol{U}$ are orthogonal Eigen vectors of $\boldsymbol{AA^T}$. The columns of $\boldsymbol{V}$ are orthogonal eigenvectors of $\boldsymbol{A^TA}$. Eigen values $\lambda_1 \ldots \lambda_r$ of $\boldsymbol{AA^T}$ are the square root of the Eigen values of $\boldsymbol{A^TA}$. The $\boldsymbol{V}$ matrix refers to terms and $\boldsymbol{U}$ matrix refers to documents.

B. Support Vector Machines:

Support Vector Machines [15] is a class of machine learning based on statistical learning and has shown excellent performance in practice. SVM addresses the general problems of learning by analyzing a linear decision boundary to discriminate between positive and negative members of a given class of high dimensional vectors. The basic idea of applying SVM can be stated by mapping the *n*-dimensional vectors into a feature space which is relevant with the selection of the kernel function. Then, SVM constructs a hyper-plane to fit into the linear or non-linear curve which then separates these two classes of vectors with the maximum margin of separation.

A maximum margin or an optimal hyper-plane is needed to lead maximal generalization when predicting classification of unlabeled example. The samples of training set used in this method consists of input vectors and they are shown as follows:

$x_i \in R_d$ (i = 1, .., n)

$y_i \in \{+1, -1\}$ (i = 1, .., n)

where, R_d refers to input space, and +1, -1 are used to stand respectively for the two classes. The analyses from the input consist of two-category target variables with two predictor variables, a linear classification rule will be used by the pair and is shown as follows:

$f(x) = (w . x_i) + b$

where, x is classified as positive if f(x) > 0 and f(x) < 0 for negative data. The decision boundary is a hyper plane:

$x_i \in R_d: (w . xi) + b = 0$

The entire process of classifier model is depicted in Figure 2. The model has been inspired from the yet another dataset classification, protein substring sequencing [16].

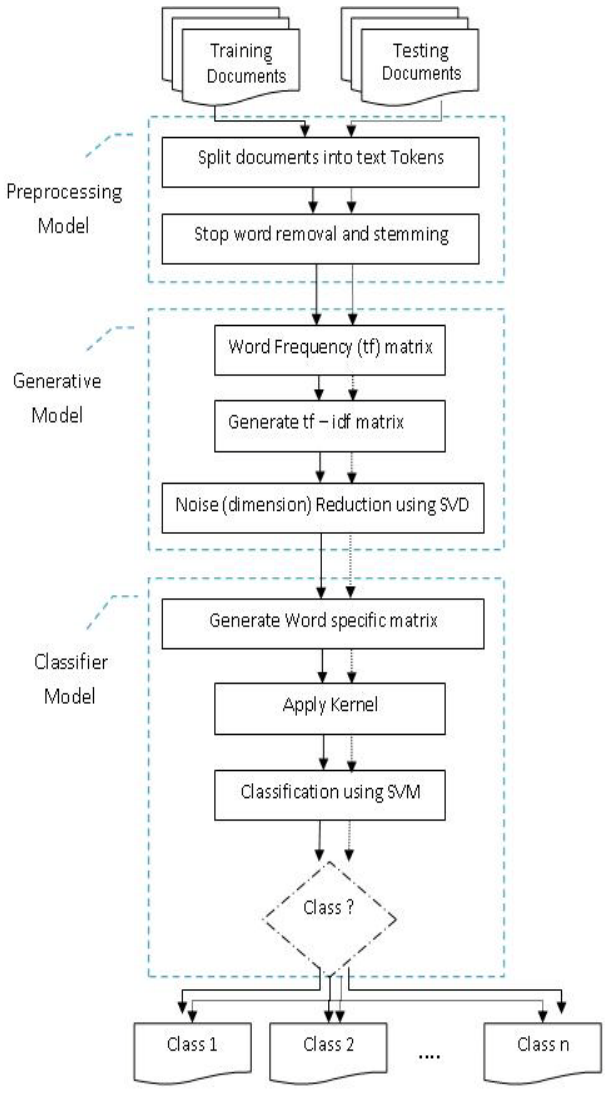

Figure 2: Document Classifier Model using SVD and SVM

Support vector machines are trained with feature vectors from the SVD output. An SVM is trained for each pairing of topics so that a classifier to distinguish between each binary pairing that is available. Linear SVMs are utilized becausetext categorization is almost always linearly separable.

The testing condition evaluates a test feature vector against each of the binary classifiers, with each classifier voting for the class that is predicted by that individual SVM. The final class is selected by choosing the vector that has the greatest number from the classifiers.

The evaluation of has been carried out by using standard measures of recall, precision and F1measure.

$$\text{precision} = \frac{\text{number of correct classifications made by the system}}{\text{total number of all classifications made by the system}}$$

$$\text{recall} = \frac{\text{number of correct classifications made by the system}}{\text{total number of all categories indicated in test documents}}$$

$$F1 = 2 * \text{precision} * \text{recall} / (\text{precision} + \text{recall})$$

Table 1 reports text classification accuracy for the Reuter-21578 data set using SVD with a Rank approximation range of values for the reduced dimension. Classification accuracy using SVM classification achieves very constant results for reduced dimensions of 200 or greater.

Table 1: F1 Score of Text categorization using SVM with SVD dimension reduction on Reuter-21578 dataset

	Rank approximation of SVD		
SVM	r = 100	r = 200	r = 300
	86.6	88.5	88.6

Table 2 shows text classification accuracy with linear kernels C=1.0 and C= 10.0 in SVMs, with and without dimension reduction on the Reuter-21578 dataset. This table shows that the prediction results in the reduced dimension are less similar to those in the original full dimensional space. So it is observed that in the reduced space obtained dimension reduction algorithm, the classification accuracy is insensitive to the choice of the kernel. Thus, we can extend further test using different kernels.

Table 2: F1 score of Text categorization using Linear kernel SVM with SVD and without dimension reduction

Kernel	**Full**	**SVD**
Linear (C = 1.0)	87.1	84.6
Linear (C = 10.0)	87.9	87.1

IV. CONCLUSION

In this paper, we reduced the dimension using SVD of document data and compared the classification using SVM with and without dimension reduction. We tested the effectiveness in classification with dimension reduction using SVM classification method. The result shows high precision accuracy, which is essentially the same as in the original full space, even with dimension reduction. They justify dimension reduction as a worthwhile preprocessing stage for achieving high efficiency and effectiveness. In SVM the indirect notation of conceptual words gave the impact in classification of documents.

V. REFERENCES

[1]. Weiguo Fan, Linda Wallace, Stephanie Rich and Zhongju Zhang, "Tapping into the Power of Text Mining", Journal of ACM, 2005, Blacksburg.

[2]. Yiming Yang and Jan O. Pedersen, A Comparative Study on Feature Selection in Text Categorization", Proceedings of the Fourteenth International Conference on Machine Learning 1997, pp. 412 – 420, ISBN:1-55860-486-3.

[3]. T. Joachims, "A Probabilistic Analysis of the Rocchio Algorithm with TFIDF for Text Categorization" Proceeding of 14th International Conference of Machine Learning, 1997, pp.143-151.

[4]. K. Aurangzeb, B. Baharum, H. L. Lam and K. Khairullah, "A Review of Machine Learning Algorithms for Text-Documents Classification", Journal of Advances in Information Technology, 2010, vol.1, pp.4 – 20.

[5]. D. L. Donoho, "High-dimensional data analysis: The curses and blessings of dimensionality", Conference of the American Mathematical Society, 2000.

[6]. L. Breiman, "Consistency for a Simple Model of Random Forests", Technical Report, Department of Statistics, University Of California, Berkeley, 2004.

[7]. D. Malathi, Dr. S. Valarmathy "A Comprehensive Survey on Dimension Reduction Techniques for Concept Extraction from a Large Corpus", IJCIS, International Journal of Computing Information Systems Vol. 3 No. 5 December 2011, Pp: 1 -6, ISSN:2229-5208

[8]. Jung-Yi Jiang, Ren-Jia Liou, and Shie-Jue Lee "A Fuzzy Self-Constructing Feature Clustering Algorithm for Text Classification", Ieee Transactions on Knowledge and Data Engineering, vol. 23, no. 3, pp 335-349 March 2011.

[9]. Laurens van der Maaten, Eric Postma, and Jaap van den Herik. "Dimensionality reduction: A comparative review", Technical Report TiCC-TR 2009-005, Tilburg University, 2009.

[10]. http://daviddlewis.com/resources/ testcollections

[11]. Stop words list, http://www.indiana.edu/cgi-bin-local/doIsearch.pl?Stopwords.

[12]. M.F. Porter, An Algorithm for Suffix Stripping, Program 14 (3) (1980) 130–137.

[13]. X.J, and W.Xo, Document Clustering with prior knowledge in Proc of ACM/SIGIR conference research and development Information Retrieval, 2006

[14]. Crammer, K. & Singer, Y., On the algorithmic implementation of multiclass kernel-based vector machines. Journal of Machine Learning Research, 2, pp. 265–292, 200.

[15]. Vapnic. 1995. The Nature of Statistical Learning Theory. Springer, New York, NY

[16]. Surayati Ismail, Razib M. Othman, Shahreen Kasim, Rohayanti Hassan, Hishammuddin Asmuni and Jumail Taliba, Pairwise Protein Substring Alignment with Latent Semantic Analysis and Support Vector Machines to Detect Remote Protein Homology, International Journal of Bio-Science and Bio-Technology, Vol.3, No. 3, September, 2011.

Short Bio Data for the Authors

D. Malathi has completed B.Sc. (Mathematics) degree and M.C.A. degree from Bharathiar University, Coimbatore in April 1998 and May 2001 respectively. She then completed M.Phil. at Bharathidasan University, Thiruchirapally, in the area of Data Structures and Algorithms in 2004. She is currently doing part time Ph.D. (Data Mining) in Anna University of Technology, Coimbatore. She has 10 years of teaching experience and currently working as Assistant Professor in the department of Computer Applications, Bannari Amman Institute of Technology, Sathyamangalam. Her research interest includes, Information Retrieval, Soft Computing and Patter Recognition. She is the life member in Indian Society for Technical Education and Member in Institution of Engineers (ISTE) and member of Computer Society of India (CSI). She has presented 3 papers in National and International Conferences and 1 paper in International Journal.

Dr. S. Valarmathy received her B.E. (Electronics and Communication Engineering) degree and M.E. (Applied Electronics) degree from Bharathiar University, Coimbatore in April1989 and January 2000 respectively. She received her Ph.D. degree at Anna University, Chennai in the area of Biometrics in 2009. She is presently working as Professor& Head in the department of Electronics and Communication Engineering, Bannari Amman Institute of Technology, Sathyamangalam. She is having a total of 20 years of teaching experience in various engineering colleges. Her research interest includes Biometrics, Image Processing, Soft Computing, Pattern Recognition and Neural Networks. She is the life member in Indian Society for Technical Education and Member in Institution of Engineers. She has published 10 papers in International and National Journals, 33 papers in International conferences and National Conferences.