ISSN (Print) : 2319-5940
ISSN (Online) : 2278-1021

International Journal of Advanced Research in Computer and Communication Engineering
Vol. 2, Issue 4, April 2013

Divisive Clustering method using Naive Bayes Algorithm for Text Categorization

K.K.Sureshkumar[1], M.Umadevi[2], Dr.N.M.Elango[3]

Assistant Professor, Department of Computer Science (P.G), Kongu Arts and Science College, Erode, India[1]

Research Scholar, Department of Computer Science (P.G), Kongu Arts and Science College, Erode, India[2]

Professor and Head, Department of MCA, RMK Engineering College, Chennai, India [3]

Abstract: This research Divisive Clustering method using Naïve Bayes algorithm for text categorization has been developed to assigning an electronic document to one or more predefined categories or classes based on its textual context. In many information processing tasks, labels are usually expensive and the unlabeled data points are abundant. To reduce the cost on collecting labels, it is crucial to predict which unlabeled examples are the most informative, i.e., improve the classifier the most if they were labelled. Many active learning techniques have been proposed for text categorization, such as SVM Active and Transductive Experimental Design. However, most of previous approaches and researches are try to discover the discriminate structure of the data space, whereas the geometrical structure is not well respected. An agglomerative clustering algorithm has been implemented where the fixed M-dimensional static window has been replaced by a dynamic window scheme using Divisive Clustering Algorithm. Because of the independence assumption, the parameters for each attribute can be learned separately, and this greatly simplifies learning, especially when the number of attributes is large. The proposed scheme is experimented using Naive Bayes Algorithm with different data set to show its better effectiveness of text categorization in terms of minimum search time. The above mentioned algorithm has been implemented using Microsoft Visual Studio .NET 2008. The coding language used is C# .NET and the back end is MS SQL Server 2005.

Keywords: Text Mining, Automatic Text Categorization (ATC), Adaptive Active Learning Algorithm, Naïve Bayes Algorithm, Divisive Clustering Algorithm

I. INTRODUCTION

Automatic E-mail has become ubiquitous in the world today. It is the long distance communication standard for millions of individuals in business, academia, and personal affairs. Its benefits are numerous, but email comes with several caveats as well. Spam mail assaults e-mail boxes daily, eating valuable server space. E-mails are sent in such great quantity that it can be frustrating and difficult to search through new mail, much less dig for an old message that is needed again. Folders and message rules, detailed mail searches and even opening separate accounts are all necessary tools today for managing mail.

Fortunately, a developing technology with many applications might help to sift through the chaos that is e-mail, and could revolutionize the way that all digital text is searched. This technology can be described as artificial intelligence text classification. Suppose a computer is given a list of categories and then shown a paragraph of text. E-mail inboxes could sort themselves, dynamic targeted advertising would improve, and web searches could display truly relevant information. If successful, the product of this research would be an automated computer program that can categorize text.

Automatic Text Categorization (ATC) scheme is the task of assigning a text document to one or more predefined categories or classes, based on its textual context. It corresponds to a supervised process, where categories are predefined by some external mechanism by establishing, at the same time, a set of already labelled examples that form the training set. Classifiers are generated from those training examples, by induction, in the so-called learning phase. This forms the machine learning paradigm (as opposed to the knowledge engineering approach) over ATC that is predominant since the 1990s exponential universalization of electronic textual information. It is further generally assumed that categories are exclusive meaning that a document can only belong to a single category (single-label

ISSN (Print) : 2319-5940
ISSN (Online) : 2278-1021

International Journal of Advanced Research in Computer and Communication Engineering
Vol. 2, Issue 4, April 2013

categorization), as this scenario has been shown to be more general than the multilabel case.

The remainder of the paper is organized as follows. Section II reviews the background study of the text categorization using different algorithms in data mining. Section III focuses on the objectives of the research. Section IV discusses about the research methodology of representing the Naïve Bayes Algorithm and Divisive Clustering Algorithm. Section V discusses about the implementation of the text level categorization method. The research findings and outcome has reveled in Section VI. Finally, conclusions were presented in Section VII.

Every channel introduces some degree of undesirable effects such as attenuation, noise, interference, and distortion. The receiver/decoder processes the received messages in order to deliver it to destination. A communication system has the basic function of transferring information (i.e., a message) from a source to a destination as shown in Fig.1.

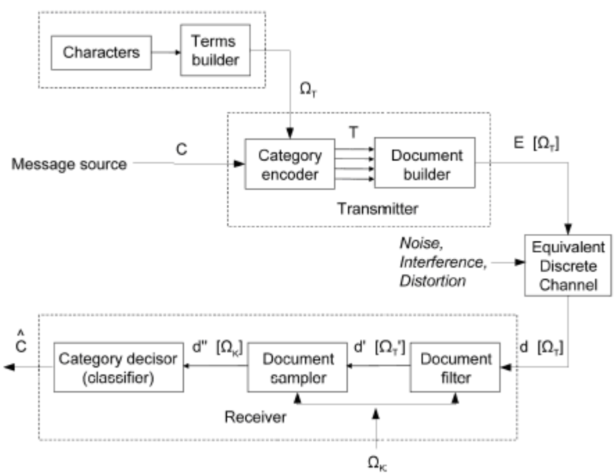


Fig.1. Model of ATC communication system.

There are mainly three essential parts of any communication system: the encoder/transmitter, the transmission channel, and the receiver/decoder. The encoder/transmitter processes the source message into the encoded and transmitted messages. The main approach in this research has been to adopt an agglomerative term clustering approach, disregarding efficiency aspects apparently improved by divisive clustering methods. A further agglomerative clustering algorithm has been implemented where the fixed M-dimensional Static window has been replaced by a dynamic window scheme. Because of the independence assumption, the parameters for each attribute can be learned separately, and this greatly simplifies learning, especially when the number of attributes is large.

II. BACKGROUND STUDY

Text classification systems already exist, but can require hours of manual training as a human teaches their computer how to recognize different types of text. If a text classification system could be designed to train itself, then would be easier to deploy and would become more viable for widespread use. Another advantage of a self-training system is that it could keep itself up-to-date on current events, which is a tedious process under a manually trained system.

High dimensionality of text can be a deterrent in applying complex learners such as Support Vector Machines (SVM) to the task of text classification. Feature clustering is a powerful alternative to feature selection for reducing the dimensionality of text data. Existing techniques for such "distributional clustering" of words are agglomerative in nature and result in (i) sub-optimal word clusters and (ii) high computational cost. In order to explicitly capture the optimality of word clusters in an information theoretic framework, first derive a global criterion for feature clustering, and then present a fast, divisive algorithm that monotonically decreases this objective function value [1, 2].

A. Divisive Information-Theoretic Feature Clustering Algorithm for Text Classification

Divisive clustering is a method of cluster analysis in which the algorithm is run repeatedly to divide clusters into sub clusters until a specified stopping point is reached. In comparison to the previously proposed agglomerative strategies divisive algorithm is much faster and achieves comparable or higher classification accuracies. Further showing that feature clustering is an effective technique for building smaller class models in hierarchical classification [5].

Agglomerative Information Bottleneck (AIB) is strictly agglomerative in nature resulting in high computational cost. Thus, AIB first selects M features (M is generally much smaller than the total vocabulary size) and then runs an agglomerative algorithm until *k* clusters are obtained (*k*_M). In order to reduce computational complexity so that it is feasible to run on the full feature set, ADC uses an alternate strategy. Agglomerative Distributional Clustering (ADC) uses the entire vocabulary but maintains only *k* word clusters at any instant. A merge of two of these clusters results in $k-1$ clusters after which a singleton cluster is created to get back *k* clusters. First, derive a global objective function to capture the decrease in mutual information due to clustering. Then present a divisive algorithm that directly minimizes this objective function, converging to a local minimum [3, 6]. Finally, providing an empirical validation of the effectiveness of the word clustering.

ISSN (Print) : 2319-5940
ISSN (Online) : 2278-1021

International Journal of Advanced Research in Computer and Communication Engineering
Vol. 2, Issue 4, April 2013

B. A Comparison of Event Models for Naive Bayes Text Classification

Recent approaches to text classification have used two different first-order probabilistic models for classification, both of which make the naive Bayes assumption. Some use a multi-variant Bernoulli model, that is, a Bayesian Network with no dependencies between words and binary word features [12]. Others use a multinomial model, that is, a unigram language model with integer word counts. This research aims to clarify the confusion by describing the differences and details of these two models, and by empirically comparing their classification performance on five text corpora [9, 15]. The multi-variant Bernoulli performs well with small vocabulary sizes, but that the multinomial performs usually performs even better at larger vocabulary sizes providing on average a 27% reduction in error over the multi-variant Bernoulli model at any vocabulary size.

Expectation-Maximization

The Expectation-Maximization (EM) algorithm is a general technique for maximum likelihood or maximum a posteriori estimation in incomplete data problems. In this task, the class labels of the unlabeled documents are considered as the missing values. The document collection D now consists of the disjoint subsets of the labelled and the unlabeled documents: The probability function of all the documents becomes:

$$P(D|\Theta) = \prod_{d_i \in D_l} P(y_i = c_j|\Theta) P(d_i|y_i = c_j; \Theta) \times \prod_{d_i \in D_u} \sum_{j=1}^{M} P(c_j|\Theta) P(d_i|c_j; \Theta) \ .$$

Because of the independence assumption, the parameters for each attribute can be learned separately, and this greatly simplifies learning, especially when the number of attributes is large [10].

C. Text Categorization with Many Redundant Features

Text categorization algorithms usually represent documents as bags of words and consequently have to deal with huge numbers of features. Most previous studies found that the majority of these features are relevant for classification, and that the performance of text categorization with support vector machines peaks when no feature selection is performed. Describing a class of text categorization problems that are characterized with many redundant features. Even though most of these features are relevant, the underlying concepts can be concisely captured using only a few features, while keeping all of them has substantially detrimental effect on categorization accuracy [7, 8].

The goal in the machine learning approach to text categorization is to devise a learning algorithm that can generate a classifier capable of categorizing (or classifying) text documents according to a number of predefined categories (or classes). This task has been mostly considered within a supervised learning scheme, but it can also be considered within an unsupervised and semi-supervised learning setups. This research work focuses on the more common supervised learning approach to text categorization [17, 11, 19].

D. A Text Clustering Framework for Information Retrieval

Text-mining methods have become a key feature for Homeland-security technologies [4, 16], as can help explore effectively increasing masses of digital documents in the search for relevant information. This research presents a model for document clustering that arranges unstructured documents into content-based homogeneous groups. The overall paradigm is hybrid because it combines pattern-recognition grouping algorithms with semantic driven processing [13]. First, a semantic-based metric measures distances between documents, by combining content-based and behavioural analysis. Such a metric allows taking into account the lexical properties, the structure and the styles characterizing the processed documents. In a second step, the model relies on a Radial Basis Function (RBF) kernel-based mapping for clustering documents. As a result, the major novelty aspect of the proposed approach is to exploit the implicit mapping of RBF kernel functions to tackle the crucial task of normalizing similarities, while embedding semantic information in the whole mechanism [14].

III. OBJECTIVES OF THE RESEARCH

The automated categorization (or classification) of texts into predefined categories has witnessed a booming interest in the last 10 years, due to the increased availability of documents in digital form and the ensuing need to organize them. In the research community the dominant approach to this problem is based on machine learning techniques: a general inductive process automatically builds a classifier by learning, from a set of pre classified documents, the characteristics of the categories. The advantages of this approach over the knowledge engineering approach (consisting in the manual definition of a classifier by domain experts) are a very good effectiveness, considerable savings in terms of expert labor power, and straightforward portability to different domains. This survey discusses the main approaches to text categorization that fall within the machine learning paradigm.

Filtering can be seen as a case of single-label TC, that is, the classification of incoming documents into two disjoint categories, the relevant and the irrelevant. Additionally, a filtering system may also further classify the documents

ISSN (Print) : 2319-5940
ISSN (Online) : 2278-1021

International Journal of Advanced Research in Computer and Communication Engineering
Vol. 2, Issue 4, April 2013

deemed relevant to the consumer into thematic categories. Similarly, an e-mail filter might be trained to discard "junk" mail and further classify non junk mail into topical categories of interest to the user.

IV. Research Methodology

Text categorization algorithms usually represent documents as bags of words and consequently have to deal with huge numbers of features. Most previous studies found that the majority of these features are relevant for classification, and that the performance of text categorization with support vector machines peaks when no feature selection is performed. Even though most of these features are relevant, the underlying concepts can be concisely captured using only a few features. Text categorization deals with assigning category labels to natural language documents. The absolute majority of works in the field employ the so-called "bag of words" approach and use plain language words as features. Using a bag of words usually leads to an explosion in the number of features, so that even moderately-sized test collections often have thousands or even tens of thousands of features. In such high-dimensional spaces, feature selection (FS) is often necessary to reduce noise and avoid over fitting.

A. Naive Bayes Algorithm

In this algorithm, each document di is generated by choosing a mixture component with the class prior probabilities P (cj | Ớ), and having this mixture component generate a document according to its own parameters, with distribution P (di|cj; Ớ). Thus, it can be written as:

The naive Bayes uses the maximum a posteriori (MAP) estimate for learning a classifier. It assumes that the occurrence of each word in a document is conditionally independent of all other words in that document given its class. Using the assumption, the probability of a document given its class becomes:

$$P(d_i|c_j;\Theta) = P(w_{i1},\ldots,w_{i|d_i|}|c_j;\Theta) \propto \prod_{k=1}^{|d_i|} P(w_{ik}|c_j;\Theta) \ .$$

The naive Bayes classifier is the simplest of these models, in that it assumes that all attributes of the examples are independent of each other given the context of the class. This is the so-called "Naive Bayes Assumption." While this assumption is clearly false in most real-world tasks, Naive Bayes often performs classification very well. This paradox is explained by the fact that classification estimation is only a function of the sign (in binary cases) of the function estimation; the function approximation can still be poor while classification accuracy remains high.

B. Divisive Clustering Algorithm

Algorithm Divisive Clustering (P, Π, l, k, W)

Input: P is the set of distributions, $\{p(C/w_t) : 1 \leq t \leq m\}$,

Π is the set of all word priors, $\{\pi_t = p(w_t) : 1 \leq t \leq m\}$,

l is the number of document classes,

k is the number of desired clusters.

Output: W is the set of word clusters $\{W_1, W_2, \ldots, W_k\}$.

- Initialization: for every word w_t, assign w_t to W_j such that $p(c_j/w_t) = \max_i p(c_i/w_t)$. This gives l initial word clusters; if $k \geq l$ split each cluster arbitrarily into atleast k/l clusters, otherwise merge the l clusters to get k word clusters.
- For each cluster W_j, compute

$$\pi(Wj) = \sum_{wt \in Wj} \pi t \ and \ p\left(\frac{c}{Wj}\right) = \sum_{wt \in Wj} \frac{\pi t}{\pi(Wj)} p\left(\frac{c}{wt}\right)$$

- Re-compute all clusters: For each word w_t, find its new cluster index as

  $j^*(w_t) = \text{argmin}_i \ KL(p(C/w_t), p(C/W_i))$, resolving ties arbitrarily. Thus compute the new word clusters W_j, $1 \leq j \leq k$, as $W_j = p\{w_t : j^*(w_t) = j\}$.
- Stop if the change in objective function value given by (13) is "small" (say 10^{-3});
- Else go to step 2.

Note: KL denotes Kullback-Leibler (KL).

V. System Implementation

The basic concern of a communication system is to transfer information from its source to a destination some distance away. Textual documents also deal with the transmission of information. Particularly, from a text categorization system point of view, the information encoded by a document is the topic or category it belongs to. Following this initial intuition, a theoretical framework is developed where Automatic Text Categorization is studied under a communication system perspective. Under this approach, the problematic indexing feature space dimensionality reduction has been tackled by a two-level supervised scheme, implemented by a noisy terms filtering and a subsequent redundant terms compression.

A communication system has the basic function of transferring information (i.e., a message) from a source to a destination. There are mainly three essential parts of any communication system: the encoder/transmitter, the transmission channel, and the receiver/decoder. This consists of the following modules. They are,

- Document Grouping.

ISSN (Print) : 2319-5940
ISSN (Online) : 2278-1021

International Journal of Advanced Research in Computer and Communication Engineering
Vol. 2, Issue 4, April 2013

- Implementing Bag-Of-Words Approach.
- Compaction and Compression.
- Analyze and Categorize.

Document Grouping

This research work concentrates more on text level categorization. The text document will be given as input to the system and it will be processed to categorize it. The textual data from the entire source in the network will be collected here. Here, it is considered with client server based approach where server received the document from more than one client. The document received is stored in database and ready for processing. Any number of document files from any number of sources is allowed here.

Implementing Bag-Of-Words Approach

The first step toward any compact document representation is the definition of the indexing features. The indexing features, also called terms, are the minimal meaningful constitutive units (a common choice is to use words). The set of different terms that appear in the collection of training documents forms the vocabulary or alphabet of terms. Once the alphabet chosen, the text document can be represented in the terms space. In this indexing process, the sequentially or order of terms in the text is commonly lost. This is known as the bag-of-words approach.

Compaction and Compression

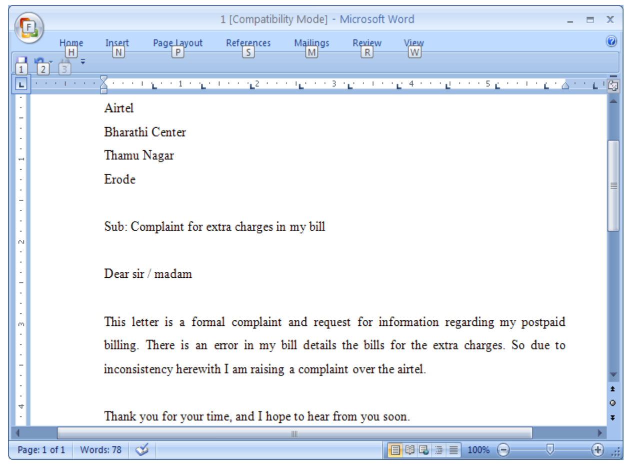
The way of compressing the text document is of two ways. It is done by compaction and compression. The compaction is done by removing the noise words. It can be done by removing the symbols like scientific or any other pneumonic symbols that cannot be used for categorization. The next way is to eliminate the redundant data like repeated words so that the unwanted scanning of words can be eliminated.

Analyze and Categorize

After eliminating all the repeated and unwanted data, the actual categorization starts. On analyzing the document with already defined categories then making them to categorize. The words in the document will be compared with the keyword in the database. If it gets matched then it will be considered as same category and it will grouped into the particular database.

VI. Research Findings

Divisive clustering algorithm outperforms various feature selection methods and previous agglomerative clustering approaches. The results can be compared with feature selection by Information Gain, Mutual Information and feature clustering using the agglomerative algorithms. Divisive Clustering achieves higher CLASSIFICATION ACCURACIES THAN THE BEST performing feature selection method. The preliminary experiments were conducted on a set of documents to test the effectiveness of this proposed system. Some set of mails are collected from the organization. Mails based on complaint, requisition, sales, acknowledgement, purchase and so on, are contained in this document set. Sample mail collected from the organization as shown in Fig 6.1.



Airtel
Bharathi Center
Thamu Nagar
Erode

Sub: Complaint for extra charges in my bill

Dear sir / madam

This letter is a formal complaint and request for information regarding my postpaid billing. There is an error in my bill details the bills for the extra charges. So due to inconsistency herewith I am raising a complaint over the airtel.

Thank you for your time, and I hope to hear from you soon.

Fig. 6.1 Sample Complaint mail categorization

The text category and keyword for the documents are presented in the Table I

TABLE I
TEXT CATEGORY AND KEYWORD FOR THE DOCUMENTS

S.No	Category	Keyword
1	Requisition	request, need, appeal
2	Complaint	complaint, defect, trouble
3	Acknowledgement	return, reply, respond
4	Inquiry	Inquiry, inquisition, analysis, inquest
5	Sales	Sales, Clearance, Selling
6	Purchase	purchase, buy, get
7	Goodwill	goodwill, beneficence
8	Feedback	feedback, suggestion, idea
9	Maths	formula, equation,
10	Computer	pen drive, cd, dvd, keyboard

Text categorization done in this proposed system is the assignment of keywords to one or more predefined categories based on their similarity to the conceptual content of the categories. The proposed system classifies the documents by the text categorization process into a set of categories namely Complaint, Requisition, Sales,

ISSN (Print) : 2319-5940
ISSN (Online) : 2278-1021

International Journal of Advanced Research in Computer and Communication Engineering
Vol. 2, Issue 4, April 2013

Acknowledgement, Purchase and Feedback. Hence the search time and the category of the document for a given particular search word is specified in Table II.

TABLE II
COMPARISON OF SEARCH TIME WITH FEATURE SELECTION METHOD AND NAÏVE BAYES ALGORITHM

S.No	Category	Keyword	Search Time with Feature Selection Method	Search Time using Naive Bayes Algorithm
1	Requisition	appeal	0.52 secs	0.16 secs
2	Complaint	trouble	0.35 secs	0.09 secs
3	Acknowledgement	return	0.36 secs	0.08 secs
4	Feedback	suggestion	0.32 secs	0.05 secs
5	Sales	clearance	0.39 secs	0.06 secs
6	Purchase	buy	0.28 secs	0.05 secs

The above Table II and the Fig.6.2 represent in graph shows the comparative results of text categorization with the existing method of Feature Selection Method using Agglomerative Clustering Algorithm and proposed method of Divisive Clustering using Naive Bayes Algorithm for its search time.

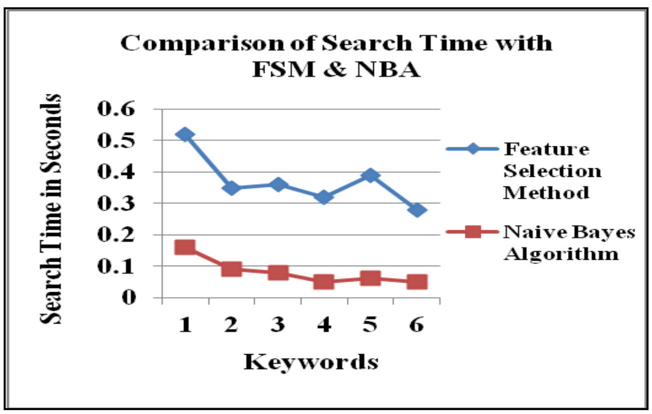


Fig. 6.2 Comparison of Search Time with FSM & NBA

It is obvious from the table that for most of the categories, proposed model shows improved search time than the existing method. Among the resultant category feedback and purchase shows the best result of 0.05 seconds. With these conclude that the Divisive Clustering method using Naive Bayes Algorithm provided an efficient and better search time for text categorization.

VII. CONCLUSION

The results show Divisive Clustering method using Naive Bayes Algorithm provided an efficient and better search time for text categorization. Divisive algorithm is much faster than the agglomerative strategies. The process may need definition of keyword and category list to perform the operation. This clustering method, with sample correlation coefficient as similarity measure, allows a high indexing term-space reduction factor with a gain of higher classification accuracy. In future work, intend to conduct experiments at a large scale on hierarchical web data to evaluate the effectiveness of the resulting hierarchical classifier. It also intend to explore local search strategies to increase the quality of the local optimal achieved by divisive clustering algorithm. Furthermore information-theoretic clustering algorithm can be applied to other applications that involve non-negative data. Future work is also foreseen in the communication theoretical modeling aspect, with special stress on the synthesis of prototype documents via the generative model proposed, as well as the deepening on the document coding (and subsequent decoding) optimal design.

ACKNOWLEDGMENT

The authors of this paper would like to express their heartfelt gratitude to the Management, Principal of their respective institutions and the faculty members of their respective departments who have contributed kind support to publish this paper. In addition, authors also wish to express their great thanks to their family members who have rendered kind support to bring this research paper for the publication. The authors also express their sincere thanks to anonymous reviewer of IJARCCE for reviewing this work and providing necessary comments.

REFERENCES

[1] Baker L.D., McCallum.A.K. (1998), "Distributional Clustering Of Words for Text Classification".
[2] Bekkerman.R, El-Yaniv.R., Tishby, N. and Winter, Y. (2003), "Distributional Word Clusters vs. Words for Text Categorization".
[3] Cataltepe, Z. and Aygun, E. (2007), "An Improvement of Centroid-Based Classification Algorithm".
[4] Chen.H, Chung.W, Xu .J.J, Wang.G, Qin.Y, Chau.M. "Crime data mining: a general framework and some examples", *IEEE Trans. Computer 37*, pp.50-56, 2004.
[5] Dhillon, L., Mallela, S. and Kumar, R. (2003),"A Divisive Information- Theoretic Feature Clustering Algorithm for Text Classification".
[6] Doan, A., Domingos, P. and Halevy, A. (2003) "Learning to match the schemas of Data Sources: A Multistrategy Approach".
[7] Joachims.T, "Text Categorization with Support Vector Machines: Learning with Many Relevant Features," *Proc. European Conf. Machine Learning (ICML)*, pp. 137-142, 1998.
[8] Joachims.T, "Transductive Inference for Text Classification Using Support Vector Machines," *Proc. Int'l Conf. Machine Learning (ICML)*, pp. 200-209, 1999.
[9] Jason D. M. Rennie, "Improving Multi-class Text Classification with Naive Bayes," 2001, Massachusetts Institute of Technology.

[10] Kamal Nigam, Andrew McCallum, Sebastian Thrun Tom Mitchell, "Text Classification from Labeled and Unlabeled Documents using EM ", Machine Learning Springer Netherlands, Vol. 39, Numbers 2-3/May 2000, pp. 103-134.
[11] Loeff.N, Frsyth.D, and Ramachandran.D, "Manifoldboost: Stagewise Function Approximation for Fully-, Semi- and Un-Supervised Learning," *Proc. 25th Int'l Conf. Machine Learning (ICML '08)*, 2005.
[12] McCallum.A and Nigam.K, "A Comparison of Event Models for Naive Bayes Text Classification". In Proceedings of AAAI-98 Workshop on Learning for Text Categorization, 1998.
[13] Mena.J, "Investigative Data Mining for Security and Criminal Detection", Butterworth-Heinemann, 2003.
[14] Scholkopf.B and Smola.A.J, "Learning with Kernels", MIT Press, 2002.
[15] Schneider.K.M, "A new feature selection score for multinomial naive bayes text classification based on KL-divergence". In Proceedings of the 42nd Meeting of the Association of Computational Linguistics (ACL), pages 186–189, 2004.
[16] Sindhwani.V, Niyogi .P, and Belkin. M, "Beyond the Point Cloud: From Transductive to Semi-Supervised Learning," *Proc. Int'l Conf. Machine Learning (ICML '05)*, 2005.
[17] Tong.S and Koller.D, "Support Vector Machine Active Learning with Application to Text Classification," J. Machine Learning Research, vol. 2, pp. 45-66, 2001.

BIOGRAPHY

K.K.Sureshkumar received a Masters Degree in Computer Science from Periyar University, in the year 2000 and M.Phil Degree in Computer Science from Manonmaniam Sundaranar University, in the year 2002 respectively. Currently, He is working as an Assistant Professor in the Department of Computer Science (P.G), Kongu Arts and Science College, Erode, affiliated to Bharathiar University. He has a total teaching experience of over 12 years. He has presented 13 papers in National Conference, 3 papers in International Conference and published 4 papers in International Journals. His research area covers Data Mining, Concurrent Engineering and currently focuses on Artificial Neural Network.

M.Umadevi received MCA Degree from Anna University, Coimbatore in 2010. Currently she is a full time M.Phil research scholar in the Department of Computer Science (P.G), Kongu Arts and Science College, Erode, affiliated to Bharathiar University. In research, currently she focuses on Data Mining especially in Text Categorization.

Dr.N.M.Elango received M.Sc. Degree and M.Phil Degree in Mathematics from Madurai Kamaraj University, Tamilnadu, India in 1983 and 1991 respectively. He received Master of Computer Applications from Manonmaniam Sundaranar University, in the year 1998 and Ph.D. Degree in Computer Science from Sastra University, Tamilnadu, India in the year 2000. Currently, He is currently Professor and Head, Department of MCA, R.M.K. Engineering College, Affiliated to Anna University, Chennai. He has a total experience of over 26 years. He has published many papers in National and International Conferences and 7 papers in Journal. His areas of research interest include image processing and Texture Component of Images.