



Text categorization via generalized discriminant analysis

Tao Li^{a,*}, Shenghuo Zhu^b, Mitsunori Ogihara^c

^a School of Computer Science, Florida International University, 11200 SW 8th Street, Miami, FL, 33199, United States

^b Internet Software Department, NEC-Labs America Inc., Cupertino, CA 95014, United States

^c Department of Computer Science, University of Rochester, Rochester, NY 14620, United States

ARTICLE INFO

Article history:

Received 30 September 2007

Received in revised form 3 February 2008

Accepted 18 March 2008

Available online 9 June 2008

Keywords:

Multi-class text categorization

GSVD

Discriminant analysis

ABSTRACT

Text categorization is an important research area and has been receiving much attention due to the growth of the on-line information and of Internet. Automated text categorization is generally cast as a multi-class classification problem. Much of previous work focused on binary document classification problems. Support vector machines (SVMs) excel in binary classification, but the elegant theory behind large-margin hyperplane cannot be easily extended to multi-class text classification. In addition, the training time and scaling are also important concerns. On the other hand, other techniques naturally extensible to handle multi-class classification are generally not as accurate as SVM. This paper presents a simple and efficient solution to multi-class text categorization. Classification problems are first formulated as optimization via discriminant analysis. Text categorization is then cast as the problem of finding coordinate transformations that reflects the inherent similarity from the data. While most of the previous approaches decompose a multi-class classification problem into multiple independent binary classification tasks, the proposed approach enables direct multi-class classification. By using generalized singular value decomposition (GSVD), a coordinate transformation that reflects the inherent class structure indicated by the generalized singular values is identified. Extensive experiments demonstrate the efficiency and effectiveness of the proposed approach.

© 2008 Published by Elsevier Ltd.

1. Introduction

With the ever-increasing growth of the on-line information and the permeation of Internet into daily life, methods that assist users in organizing large volumes of documents are in huge demand. In particular, automatic text categorization has been extensively studied recently. This categorization problem is usually viewed as supervised learning, where the goal is to assign predefined category labels to unlabeled documents based on the likelihood inferred from the training set of labeled documents. Numerous approaches have been applied, including Bayesian probabilistic approaches (Lewis, 1998; Tzeras & Hartmann, 1993), nearest neighbor (Lam & Ho, 1998; Masand, Linoff, & Waltz, 1992), neural networks (Wiener, Pedersen, & Weigend, 1995), decision trees (Apte, Damerau, & Weiss, 1998), inductive rule learning (Cohen & Singer, 1996; Dumais, Platt, Heckerman, & Sahami, 1998), support vector machines (Godbole, Sarawagi, & Chakrabarti, 2002; Joachims, 2001), maximum entropy (Nigam, Lafferty, & McCallum, 1999), boosting (Schapire & Singer, 2000), and linear discriminate projection (Chakrabarti, Roy, & Soundalgekar, 2002) (see (Yang & Liu, 1999) for comparative studies of text categorization methods).

Although document collections are likely to contain many different categories, most of the previous work was focused on binary document classification. One of the most effective binary classification techniques is the support vector machines (SVMs) (Vapnik, 1998). It has been demonstrated that the method performs superbly in binary discriminative text

* Corresponding author.

E-mail address: taoli@cs.fiu.edu (T. Li).

classification (Joachims, 2001; Yang & Liu, 1999). SVMs are accurate and robust, and can quickly adapt to test instances. However, the elegant theory behind the use of large-margin hyperplanes cannot be easily extended to multi-class text categorization problems. A number of techniques for reducing multi-class problems to binary problems have been proposed, including one-versus-the-rest method, pairwise comparison (Hastie & Tibshirani, 1998) and error-correcting output coding (Allwein, Schapire, & Singer, 2000; Dietterich & Bakiri, 1995). In these approaches, the original problems are decomposed into a collection of binary problems, where the assertions of the binary classifiers are integrated to produce the final output. In practice, which reduction method is best suited is problem-dependent, so it is a non-trivial task to select the decomposition method. Indeed, each reduction method has its own merits and limitations (Allwein et al., 2000). In addition, regardless of specific details, these reduction techniques do not appear to be well suited for text categorization tasks with a large number of categories, because training of a single, binary SVM requires $O(n^2)$ time for $1.7 \leq \alpha \leq 2.1$, where n is the number of training data (Joachims, 1998). Thus, having to train many classifiers has a significant impact on the overall training time. Also, the use of multiple classifiers slows down prediction. Thus, despite its elegance and superiority, the use of SVM may not be best suited for multi-class document classification. However, there do not appear to exist many alternatives, since many other techniques that can be naturally extended to handle multi-class classification problems, such as neural networks and decision trees, are not so accurate as SVMs (Yang & Liu, 1999; Yang & Pederson, 1997).

In statistics pattern recognition literature, discriminant analysis approaches are well-known to be able to learn discriminative feature transformations (see, e.g., (Fukunaga, 1990)). For example, Fisher discriminant analysis (Fisher, 1936) finds a discriminative feature transformation as eigenvectors associated with the largest eigenvalues of matrix $T = S_w^{-1}S_b$, where S_w is the intra-class covariance matrix and S_b is the inter-class covariance matrix.¹ Intuitively, T captures not only compactness of individual classes but separations among them. Thus, eigenvectors corresponding to the largest eigenvalues of T are likely to constitute a discriminative feature transform. However, for text categorization, S_w is usually singular owing to the large number of terms. Simply removing the null space of S_w would eliminate important discriminant information when the projections of S_b along those directions are not zeros (Fukunaga, 1990). This issue has stymied attempts to use traditional discriminant approaches in document analysis.

In this paper we resolve this problem. We extend discriminant analysis and present a simple, efficient, but effective solution to text categorization. We cast text categorization as the problem of finding transformations to reflect the inherent similarity from the data. In this framework, given a document of unknown class membership, we compare the distance of the new document to the centroid of each category in the transformed space and assign it to the class having the smallest distance to it. We call this method generalized discriminant analysis (GDA), since it uses generalized singular value decomposition to optimize transformation. We show that the transformation derived using GDA is equivalent to optimization via the determinant ratios and a new criterion. A preliminary version of the work has been presented in a conference paper (Li, Zhu, & Ogihara, 2003).

GDA has several favorable properties: first, it is simple and can be programmed in a few lines in MATLAB. Second, it is efficient. (Most of our experiments only took several seconds.) Third, the algorithm does not involve parameter tuning. Finally, and probably the most importantly, it is very accurate. We have conducted extensive experiments on various datasets to evaluate its performance. The rest of the paper is organized as follows: Section 2 reviews the related work on text categorization. Section 3 introduces classical linear discriminant analysis. Section 4 presents the generalized discriminant analysis for handling singular problems. Section 5 shows that the transformation of derived using GDA can also be obtained by optimizing determinant ratios and a new criterion. Section 6 presents some illustrating examples. Section 7 shows experimental results. Finally, Section 8 provides conclusions and discussions.

2. Related work

Text categorization algorithms can be roughly classified into two types: those algorithms that can be naturally extended to handle multi-class cases and those require decomposition into binary classification problems. The first consists of such algorithms as Naive Bayes (Lam & Ho, 1998; Masand et al., 1992), Neural Networks (Ng, Goh, & Low, 1997; Wiener et al., 1995), K-Nearest Neighbors (Lam & Ho, 1998; Masand et al., 1992), Maximum Entropy (Nigam et al., 1999) and decision trees. Naive Bayes uses the joint distributions of words and categorizes to estimate the probabilities that an input document belongs to each document class and then selects the most probable class. K-Nearest Neighbor finds the k nearest neighbors among training documents and uses the categories of the k neighbors to determine the category of the test document. The underlying principle of maximum entropy is that without external knowledge, uniform distribution should be preferred. Based on this principle, it estimates the conditional distribution of the class label given a document.

The reduction techniques that are used by the second group include one-versus-the-rest method (Scholkopf & Smola, 2002), error-correcting output coding (Dietterich & Bakiri, 1995), pairwise comparison (Hastie & Tibshirani, 1998), and multi-class objective functions, where the first two have been applied to text categorization (Ghani, 2000; Yang & Liu, 1999).

In the one-versus-the-rest method a classifier separating between from a class and the rest is trained for each class. Multi-class classification is carried out by integrating prediction of these individual classifiers with a strategy for resolving

¹ This is equivalent to using eigenvectors associated with the smallest eigenvalues of matrix $T = S_b^{-1}S_w$. It indicates that traditional discriminant analysis requires the non-singularity of at least one covariance matrix. Since the rank of S_w is usually greater than that of S_b , we will base our discussion on the eigenvalue-decomposition of $T = S_w^{-1}S_b$.

conflicts. The method is sometimes criticized for solving asymmetric problems in a symmetrical manner and for not considering correlations between classes.

Error-correcting output coding (ECOC) (Dietterich & Bakiri, 1995) partitions the original set of classes into two sets in many different ways. A binary classifier is trained for each partition. The partitions are carefully chosen so that the outputs of these classifiers assign a unique binary codeword for each class (with a large Hamming distance between any pair of them). The class of an input with unknown class membership is chosen by computing the outputs of the classifiers on that input and then finding the class with the codeword closest to the output codeword.

Although SVMs are considered to be very effective in binary classification, its large training costs may make it unsuitable for multi-class classification with a large number of classes if the above decomposition techniques are applied. Also, the lack of a clear winner among the above techniques makes the reduction task complicated. Our GDA directly deals with multi-class classification and does not require reduction to binary classification problems.

Other techniques for text categorization exist. Godbole et al. (2002) propose a new multi-class classification technique that exploits the accuracy of SVMs and the speed of Naive Bayes. It uses a Naive Bayes classifier to compute a confusion matrix quickly. Then it uses this matrix to reduce both the number and the complexity of binary SVMs to be built. Chakrabarti et al. (2002) propose a fast text classification technique that uses multiple linear projections. It first projects training instances to low-dimensional space and then builds decision tree classifiers on the projected spaces. Fragoudis, Meretakos, and Likothanassis (2002) propose a new algorithm that targets both feature and instance selection for text categorization.

In summary, as pointed out in Yang and Liu (1999); Scholkopf and Smola (2002), it is fair to say that there is probably no multi-class approach generally outperforms the others. For practical problems, the choice of approach will depend on constraints on hand such as required accuracy, the time available for development and training and the nature of the classification problem. The simple, efficient and accurate generalized discriminant analysis provides a good choice for text multi-class classification.

3. Classical linear discriminant analysis

The notations used through the discussion of this paper are listed in Table 1.

Given a document-term matrix $A = (a_{ij}) \in \mathfrak{R}^{n \times N}$, where each row corresponds to a document and each column corresponds to a particular term, we consider finding a linear transformation $G \in \mathfrak{R}^{N \times \ell}$ ($\ell < N$) that maps each row a_i ($1 \leq i \leq n$) of A in the N -dimensional space to a row y_i in the ℓ -dimensional space. The resulting data matrix $A^\ell = AG \in \mathfrak{R}^{n \times \ell}$ contains ℓ columns, i.e., there are ℓ features for each document in the reduced (transformed) space. It is also clear that the features in the reduced space are linear combinations of the features in the original high dimensional space, where the coefficients of the linear combinations depend on the transformation G . Linear discriminant projection tries to compute the optimal transformation matrix G such that the class structure is preserved. More details are given below.

Assume there are k classes in the data set. Suppose m_i, S_i, P_i are the mean vector, covariance matrix, and a prior probability of the i th class, respectively, and m is the total mean. For the covariance matrix S_i for the i th class, we can decompose it as $S_i = X_i X_i^T$, where X_i is the square root of S_i and has the same number of columns as the number of data points in the i th class. Define the matrices

$$H_b = [\sqrt{P_1}(m_1 - m), \dots, \sqrt{P_k}(m_k - m)] \in \mathfrak{R}^{N \times k},$$

$$H_w = [\sqrt{P_1}X_1, \dots, \sqrt{P_k}X_k] \in \mathfrak{R}^{N \times n}.$$

Then the between-class scatter matrix S_b , the within-class scatter matrix S_w , and the total scatter matrix S_t are defined as follows (Fukunaga, 1990):

$$S_b = \sum_{i=1}^k P_i (m_i - m)(m_i - m)^T = H_b H_b^T,$$

Table 1
Notations

Notations	Descriptions
A	document-term matrix
n	number of data points, i.e., documents
N	number of the dimensions, i.e., terms
k	number of class
S_i	covariance matrix of the i th class
S_b	between-class scatter matrix
S_w	within-class scatter matrix
S_t	total scatter matrix
G	reduction transformation
m_i	centroid of the i th class
m	global centroid of the training set

$$S_w = \sum_{i=1}^k P_i S_i = H_w H_w^T,$$

$$S_t = S_b + S_w.$$

In the lower-dimensional space resulting from the linear transformation G , the within-cluster and between-cluster matrices become

$$S_w^L = (G^T H_w)(G^T H_w)^T = G^T S_w G,$$

$$S_b^L = (G^T H_b)(G^T H_b)^T = G^T S_b G.$$

An optimal transformation G would maximize $\text{Trace}(S_b^L)$ and minimize $\text{Trace}(S_w^L)$. A common optimization for computing optimal G is

$$G^* = \underset{G}{\text{argmax}} \text{Trace}((G^T S_w G)^{-1} G^T S_b G).$$

The solution can be readily obtained by solving a eigenvalue decomposition problem on $S_w^{-1} S_b$, provided that the within-class scatter matrix S_w is non-singular. Since the rank of the between-class scatter matrix is bounded above by $k - 1$, there are at most $k - 1$ discriminant vectors.

4. Generalized discriminant analysis

In general, the within-class scatter matrix S_w may be singular especially for document-term matrix where the dimension is very high. Usually, this problem is overcome by using a non-singular intermediate space of S_w obtained by removing the null space of S_w and then computing eigenvectors. However, the removal of the null space of S_w possibly eliminates some useful information because some of the most discriminant dimensions may be lost by the removal. In fact, the null space of S_w is guaranteed to contain useful discriminant information when the projections of S_b are not zeros along those directions. Thus, simple removal of the null space of S_w is not an effective resolution (Fukunaga, 1990). A common way to deal with it is to use generalized eigenvalue decomposition (Howland & Park, 2003; Li et al., 2003).

4.1. The basics of GSVD

Singular value decomposition (SVD) is a process of decomposing a rectangular matrix into three other matrices of a very special form. It can be viewed as a technique for deriving a set of uncorrelated indexing variables or factors (Deerwester, Dumais, Landauer, Furnas, & Harshman, 1990). A generalized singular value decomposition (GSVD) is an SVD of a sequence of matrices. GSVD has played a significant role in signal processing and in signal identification and has been widely used in such problems as source separation, stochastic realization and generalized Gauss–Markov estimation. The diagonal form of GSVD, shown below, was first introduced in Loan (1976).

Theorem 1 (GSVD diagonal form, Loan, 1976). *If $A \in \mathfrak{R}^{m \times p}$, $B \in \mathfrak{R}^{n \times p}$, and $\text{rank}(A^T, B^T) = k$, then there exist two orthogonal matrices, $U \in \mathfrak{R}^{m \times m}$ and $V \in \mathfrak{R}^{n \times n}$, and a non-singular matrix, $X \in \mathfrak{R}^{p \times p}$, such that*

$$\begin{bmatrix} U^T & 0 \\ 0 & V^T \end{bmatrix} \begin{bmatrix} A \\ B \end{bmatrix} X = [CS][I_k \ 0], \tag{1}$$

where C and S are non-negative diagonal and of dimension $m \times k$ and $n \times k$, respectively, $1 \geq S_{11} \geq \dots \geq S_{\min(n,k), \min(n,k)} \geq 0$, and $C^T C + S^T S = I_k$.

The *generalized singular values* are defined to be the component-wise ratios of the diagonal entries of the two diagonal matrices. In signal processing, A is often the signal matrix and B is the noise matrix, in which case the generalized singular values are referred to as signal–noise ratios. To compute GSVD, we use the built-in GSVD function in Matlab: basically GSVD uses the CS decomposition (Golub & Loan, 1996), as well as the SVD decomposition and QR decomposition. For more discussions on computing GSVD, please refer to (Bai et al., 1992).

4.2. Generalized discriminant analysis

Let $K = [H_b H_w]^T$, which is a $k + n$ by N matrix. By the generalized singular value decomposition, there exist orthogonal matrices $U \in \mathfrak{R}^{k \times k}$, $V \in \mathfrak{R}^{n \times n}$, and a non-singular matrix $X \in \mathfrak{R}^{N \times N}$, such that

$$\begin{bmatrix} U^T & 0 \\ 0 & V^T \end{bmatrix} KX = \begin{bmatrix} \Sigma_1 & 0 \\ 0 & 0 \\ \Sigma_2 & 0 \\ 0 & 0 \end{bmatrix}, \tag{2}$$

where

$$\begin{aligned} \Sigma_1 &= \text{diag}(\overbrace{1, \dots, 1}^r, \alpha_1, \dots, \alpha_s, \overbrace{0, \dots, 0}^{t-r-s}), \\ \Sigma_2 &= \text{diag}(\overbrace{0, \dots, 0}^r, \beta_1, \dots, \beta_s, \overbrace{1, \dots, 1}^{t-r-s}), \\ t &= \text{rank}(K), \quad r = t - \text{rank}(H_w^T), \\ s &= \text{rank}(H_b) + \text{rank}(H_w) - t, \end{aligned}$$

satisfying

$$\begin{aligned} 1 &> \alpha_1 \geq \dots \geq \alpha_s > 0, \\ 0 &< \beta_1 \leq \dots \leq \beta_s < 1, \\ \text{and } \alpha_i^2 + \beta_i^2 &= 1 \quad \text{for } i = 1, \dots, s. \end{aligned}$$

From Eq. (2), we have

$$(X^T H_b)(X^T H_b)^T = (X^T S_b X) = \begin{bmatrix} \Sigma_1^T \Sigma_1 & 0 \\ 0 & 0 \end{bmatrix}, \tag{3}$$

$$(X^T H_w)(X^T H_w)^T = (X^T S_w X) = \begin{bmatrix} \Sigma_2^T \Sigma_2 & 0 \\ 0 & 0 \end{bmatrix}. \tag{4}$$

Note that entries on the main diagonal of Σ_1 are non-increasing while entries on the main diagonal of Σ_2 is less than 1 and non-decreasing. Hence, a natural extension of the proposed linear discriminant analysis (e.g., maximizing $\text{Trace}(S_b^L)$ while minimizing $\text{Trace}(S_w^L)$) in Section 3 is to choose the first k columns of the matrix X in Eq. (2) as the transformation matrix G . The transformation matrix G can also be obtained by optimizing the determinant ratios or using a new optimization criterion as in Section 5.

4.3. The GDA algorithm

Once the transformation G has been determined, classification is performed in the transformed space based on a distance metrics, such as Euclidean distance

$$d(\mathbf{x}, \mathbf{y}) = \sqrt{\sum_i (x_i - y_i)^2}$$

and cosine measure

$$d(\mathbf{x}, \mathbf{y}) = 1 - \frac{\sum_i x_i y_i}{\sqrt{\sum_i x_i^2} \sqrt{\sum_i y_i^2}}.$$

A new instance, \mathbf{z} , it is classified to

$$\underset{k}{\text{argmin}} d(\mathbf{z}G, \bar{\mathbf{x}}_k G), \tag{5}$$

where $\bar{\mathbf{x}}_k$ is the centroid of k th class.

The pseudo codes of the training and prediction procedures are described as follows:

Algorithm 1 (Training procedure $G = \text{Train}(\mathbf{x}'s)$).

- Input: the training data x_i 's
- Output: the transformation G ;
- begin**
- 1. Construct the matrices H_w and H_b ;
- 2. Perform GSVD on the matrix pair;
- 3. Obtain G as described in Section 4.2.
- 4. **Return** G ;
- end**

Algorithm 2 (Prediction procedure $T = \text{Predict}(G, \mathbf{x})$).

- Input: the transformation G generated by the training procedure; and a new instance \mathbf{x} ;
- Output: the label T of the new instance;

begin

1. Perform Prediction as in Eq. (5);
2. **Return** T ;

end

5. Connections

Here we show that the above derivation can also be obtained by optimizing the determinant ratios or using a new optimization criterion.

5.1. Optimizing the determinant ratio

Fisher’s criterion is to maximize the ratio of the determinant of the inter-class scatter matrix of the projected samples to the determinant of the intra-class scatter matrix of the projected samples:

$$\mathcal{J}(G) = \frac{|G^T S_b G|}{|G^T S_w G|}. \tag{6}$$

One way to overcome the requirements of non-singularity of Fisher’s criterion is looking for solutions that simultaneously maximize $|G^T S_b G|$ minimize $|G^T S_w G|$. Using GSVD, H_b^T and H_w^T are decomposed as

$$H_w^T = UC[\mathbf{I}_k \mathbf{0}]X^{-1} \quad \text{and} \quad H_b^T = VS[\mathbf{I}_k \mathbf{0}]X^{-1}.$$

To maximize $\mathcal{J}(G)$, $|G^T S_b G|$ should be increased while decreasing $|G^T S_w G|$. Let $C' = C[\mathbf{I}_k \mathbf{0}]$ and $S' = S[\mathbf{I}_k \mathbf{0}]$. Then we have

$$S_b = H_b H_b^T = X S'^2 X^{-1} \quad \text{and} \quad S_w = H_w H_w^T = X C'^2 X^{-1}.$$

This implies

$$\begin{aligned} |G^T S_b G| &= |G^T X S'^2 X^{-1} G| \\ &= (|S' X^{-1} G|)^2, \\ |G^T S_w G| &= |G^T X C'^2 X^{-1} G| \\ &= (|C' X^{-1} G|)^2. \end{aligned}$$

Note that the entries on the main diagonal of S is decreasing while the entries on the main diagonal of C is decreasing. Thus, the matrix G satisfying

$$X^{-1}G = \begin{bmatrix} \mathbf{I}_k \\ \mathbf{0} \end{bmatrix}$$

would simultaneously maximize $|G^T S_b G|$ and minimize $|G^T S_w G|$. So, we have

$$G = X \begin{bmatrix} \mathbf{I}_k \\ \mathbf{0} \end{bmatrix}.$$

In the case where we must weight the transformation with the generalized singular,

$$G = X \begin{bmatrix} \mathbf{I}_k \\ \mathbf{0} \end{bmatrix} S^T$$

is the transformation we want.

5.2. A new criterion

5.2.1. The criterion

We show that the generalized discriminant analysis can also be derived using the following criterion for discriminating inter-class and intra-class distances by inter-class and intra-class covariance matrices:

$$\min_G \{ \|H_b^T G - I_n\|_F^2 + \|H_w^T G\|_F^2 \}, \tag{7}$$

where $\|X\|_F$ is the Frobenius norm of the matrix X , i.e., $\sqrt{\sum_{i,j} x_{ij}^2}$. The criterion does not involve the inverse of the intra-class matrix and is similar to Tikhonov regularization of least squares problems. Intuitively, the first term of (7) is used to minimize the difference between the projection of $\mathbf{x}_i - \mathbf{x}$ in a new space and the i th unit vector of the new space. The second term is used to minimize the intra-class covariance.

The Eq. (7) can be rewritten as

$$\min_G \left\| \begin{bmatrix} H_w^T \\ H_b^T \end{bmatrix} G - \begin{bmatrix} 0 \\ I_n \end{bmatrix} \right\|_F^2 \quad (8)$$

and this is a least squares problem with the solution

$$(H_w H_w^T + H_b H_b^T) G = H_b^T. \quad (9)$$

5.2.2. Stable solution

Here we will show how to use GSVD to compute efficiently the solution to the optimization problem and show that the solution thus obtained is stable.

By plugging the GSVD matrices of H_w and H_b , e.g.,

$$H_w^T = UC[\mathbf{I}_k \mathbf{0}]X^{-1} \quad \text{and} \quad H_b^T = VS[\mathbf{I}_k \mathbf{0}]X^{-1},$$

in (9), we have

$$G = X \begin{bmatrix} \mathbf{I}_k \\ \mathbf{0} \end{bmatrix} S^T V^T.$$

Since V is orthogonal, we can drop it without changing the squared distance. So, we have

$$G = X \begin{bmatrix} \mathbf{I}_k \\ \mathbf{0} \end{bmatrix} S^T. \quad (10)$$

This derivation of G holds even if S_w is singular. Thus, by using GSVD to solve the new criterion, we can avoid removing null space, thereby keeping all the useful information. We now state a theorem that shows that the solution is stable.

Theorem 2 (GSVD relative perturbation bound, Demmel and Veselić, 1992). *Suppose A and B be matrices with the same number of columns and B is of full column rank. Let $A = A_1 D_1$ and $B = B_1 D_2$ such that D_1 and D_2 have full rank. Let $E = E_1 D_1$ and $F = F_1 D_2$ be perturbations of A and B , respectively, such that for all x there exist some $\eta_1, \eta_2 < 1$ for which it holds that*

$$\|E_1 x\|_2 \leq \eta_1 \|A_1 x\|_2, \quad \|F_1 x\|_2 \leq \eta_2 \|B_1 x\|_2.$$

Let σ_i and $\tilde{\sigma}_i$ be the i th generalized singular value of (A, B) and that of $(A + E, B + F)$, respectively. Then either $\sigma_i = \tilde{\sigma}_i = 0$ or

$$\frac{|\sigma_i - \tilde{\sigma}_i|}{\sigma_i} \leq \frac{\eta_1 + \eta_2}{1 - \eta_2}.$$

The above theorem gives a bound on the relative error of the generalized eigenvalues (C_{ii} and S_{ii}) if the difference between the estimated covariance matrices and the genuine covariance matrices is small. This guarantees that the relative error of G is bounded by the relative error of estimated intra- and inter-class covariance matrices.

GSVD also brings some favorable features, which might improve accuracy. In particular, computation of the cross products $H_b H_b^T$ and $H_w H_w^T$, which causes roundoff errors, is not required.

6. Text classification via GDA: examples

A well-known transformation method in information retrieval is latent semantic indexing (LSI) (Deerwester et al., 1990), which applies singular value decomposition (SVD) to the document-term matrix and computes eigenvectors having largest eigenvalues as the directions related to the dominant combinations of the terms occurring in the dataset (*latent semantics*). A transformation matrix constructed from these eigenvectors projects a document onto the latent semantic space. Although LSI has been proven extremely useful in information retrieval, it is not optimal for text categorization because LSI is completely unsupervised. In other words, LSI deals with the data without paying any particular attention to the underlying class structure. It only aims at optimally transforming the original data into a lower dimensional space with respect to the mean squared error, which has nothing to do with the discrimination of the different classes. Our GDA approach possesses advantages of both discriminant analysis and of latent semantic analysis. By explicitly taking the intra-class and inter-class covariance matrices into the optimization criterion, GDA deals directly with discrimination between classes. Furthermore, by employing GSVD to solve the optimization problem, GDA tries to identify the latent concepts indicated by the generalized singular values.

To illustrate how well GDA can perform, we present here two examples. In the first example, we compare GDA against LDA and LSI. Fig. 1 shows a small dataset consisting of nine phrases in three topics: user interaction, graph theory, and distributed systems.

After removing words (terms) that occurs only once, we have the document-term matrix as shown in Fig. 2.

The first and second samples in each class are used for training. GDA, LDA, and LSI are run on the training data to obtain transformation matrices. Fig. 3 shows the plot of the distances/similarities between document pairs in the transformed space using each of the three methods.

No.	Class	Phrase
1	1	Human interface for user response
2	1	A survey of user opinion of computer system response time
3	1	Relation of user-perceived response time to error measurement
4	2	The generation of random, binary, unordered trees
5	2	The intersection graph of paths in trees
6	2	Graph Minors IV: Widths of trees and well-quasi-ordering
7	3	A survey of distributed shared memory system
8	3	RADAR: A multi-user distributed system
9	3	Management interface tools for distributed computer system

Fig. 1. Nine example sentences.

word \ No.	1	2	3	4	5	6	7	8	9
a		1					1	1	
computer		1							1
distributed							1	1	1
for	1								1
graph					1	1			
interface	1								1
of		2	1	1	1	1	1		
response	1	1	1						
survey		1					1		
system		1					1	1	1
the				1	1				
time		1	1						
trees				1	1	1			
user	1	1	1					1	

Fig. 2. Document-term matrix.

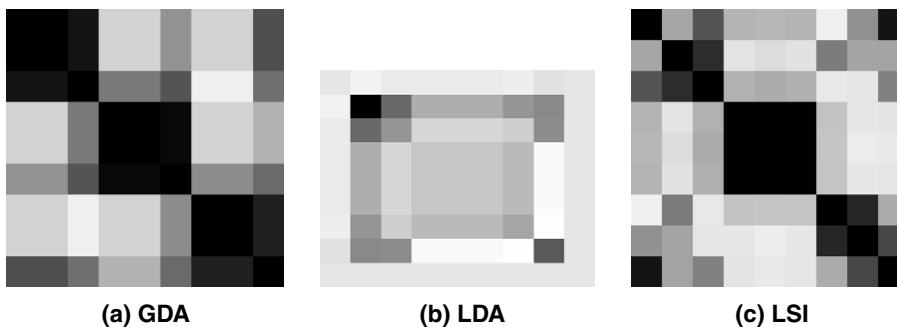


Fig. 3. Pairwise document similarity via GDA, LDA, and LSI. The darker the close is the more similar the documents are. GDA is a clear winner.

The second example illustrates differences between GDA and LSI. Distinction among three newsgroups in 20NG are attempted by selecting from each newsgroup 20 training and 20 for testing. Fig. 4 shows plots of the the 60 testing articles using the two dominant directions as the axes. GDA has clear separation while the LSI plot shows an L-shaped concentration of the data points. The confusion matrices of these methods are shown in Table 2. GDA clearly performed better than LSI.

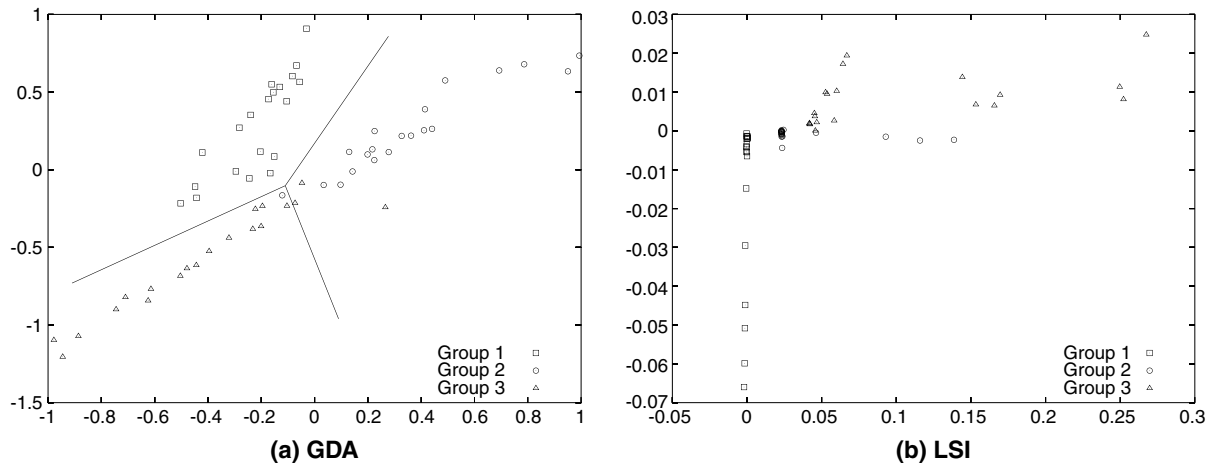


Fig. 4. Document plots. The three groups are separated significantly better with GDA than with LSI.

Table 2

The confusion matrices (left, GDA; right, LSI)

Actual	Prediction			Actual	Prediction		
	1	2	3		1	2	3
1	20	0	0	1	20	0	0
2	0	19	1	2	0	3	17
3	0	0	0	3	7	5	8

7. Experiments

7.1. The datasets

For our experiments we used a variety of datasets, most of which are frequently used in the information retrieval research. The range of the number of classes is from 4 to 105 and the range of the number of documents is from 476 to 20,000, which seem varied enough to obtain good insights as to how GDA performs. Table 3 summarizes the characteristics of the datasets.

7.1.1. 20Newsgroups

The 20Newsgroups (20NG) dataset contains approximately 20,000 articles evenly divided among 20 Usenet newsgroups. The raw text size is 26MB. All words were stemmed using a Porter stemming program (Porter, 1997), all HTML tags were skipped, and all header fields except subject and organization of the posted article were ignored.

7.1.2. WebKB

The WebKB dataset² contains Web pages collected from University Computer Science Departments. There are approximately 8300 documents in the set and they are divided into seven categories: student, faculty, staff, course, project, department, and other. The raw text size of the dataset is 27MB. Among the seven categories, student, faculty, course, and project are the four most populous. The subset consisting only of these categories is also used here, which is called WebKB4. In neither of the datasets, we used stemming or stop lists.

7.1.3. Industry sector

The Industry Section dataset³ is based on the data made available by Market Guide, Inc. (<http://www.marketguide.com>). The set consists of company homepages that are categorized in a hierarchy of industry sectors, but we disregarded the hierarchy. There were 9637 documents in the dataset, which were divided into 105 classes. We tokened the documents by skipping all MIME and HTML headers and using a standard stop list. We did not perform stemming.

² Both 20NG and WebKB are available at <http://www-2.cs.cmu.edu/afs/cs/project/theo-11/www/wwkb>.

³ Available at <http://www.cs.cmu.edu/TextLearning/datasets.html>.

Table 3
Data sets descriptions

Datasets	# Documents	# Class
20NG	20,000	20
WebKB4	4199	4
WebKB	8280	7
Industry sector	9637	105
Reuters-Top10	2900	10
Reuters-2	9000	50
CSTR	476	4
K-dataset	2340	20
TDT2	7980	96

7.1.4. Reuters

The Reuters-21578 Text Categorization Test Collection contains documents collected from the Reuters newswire in 1987. It is a standard text categorization benchmark and contains 135 categories. We used its subsets: one consisting of the ten most frequent categories, which we call Reuters-Top10, and the other consisting of documents associated with a single topic, which we call Reuters-2. Reuters-2 had approximately 9000 documents and 50 categories.

7.1.5. TDT2

TDT2 is the NIST Topic Detection and Tracking text corpus version 3.2 released in December 6, 1999 (TDT2, 1998). This corpus contains news data collected daily from nine news sources in two languages (American English and Mandarin Chinese), over a period of six months (January–June in 1998). We used only the English news texts, which were collected from New York Times Newswire Service, Associated Press Worldstream Service, Cable News Network, Voice of America, American Broadcasting Company, and Public Radio International. The documents were manually annotated using 96 target topics. We selected the documents having annotated topics and removed the brief texts. The resulting dataset contained 7980 documents.

7.1.6. K-dataset

This dataset was obtained from the WebACE project (Han et al., 1998). It contained 2340 documents consisting of news articles from Reuters News Service made available on the Web in October 1997. These documents were divided into 20 classes. They were processed by eliminating stop words and HTML tags, stemming the remaining words using Porter's suffix-stripping algorithm.

7.1.7. CSTR

This is the dataset of the abstracts of technical reports published in the Department of Computer Science at the University of Rochester between 1991 and 2002.⁴ The dataset contained 476 abstracts, which were divided into four research areas: Symbolic-AI, Spatial-AI, Systems, and Theory. We processed the abstracts by removing stop words and applying stemming operations on the remaining words.

7.2. Data preprocessing

In all experiments, we randomly chose 70% of the documents for training and assigned the rest for testing. It is suggested in Yang and Pederson (1997) that information gain is effective for term removal and it can remove up to 90% or more of the unique terms without performance degrade. So, we first selected the top 1000 words by information gain with class labels. The feature selection is done with the Rainbow package (McCallum, 1996).

Here we use classification accuracy for evaluation. Different measures, such as precision-recall graphs and F_1 measure (Yang & Liu, 1999), have been used in the literature. However, since the datasets used in our experiments are relatively balanced and single-labeled, and our goal in text categorization is to achieve low misclassification rates and high separation between different classes on a test set, we thought that accuracy is the best measure of performance. All of our experiments were carried out on a P4 2GHz machine with 512M memory running Linux 2.4.9–31.

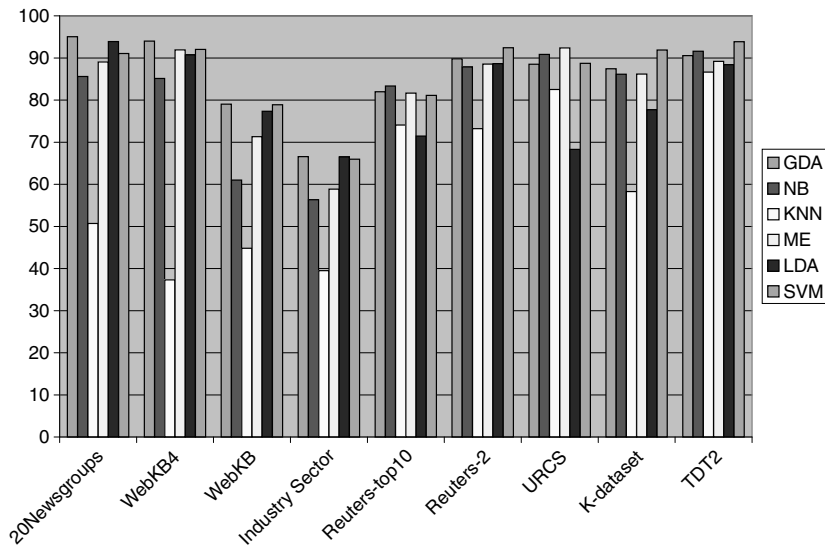
7.3. Experimental results

Now we present and discuss the experimental results. Here we compare GDA against Naive Bayes (NB for short), K-Nearest Neighbor (KNN for short), Maximum Entropy (ME for short), LDA, and SVM on the same datasets with the same training and testing data. Recall that the first three of the methods we compare against are commonly used direct methods for multi-class classification (in the sense that they do not require reduction to binary classification problems). For experiments

⁴ The TRs are available at <http://www.cs.rochester.edu/trs>.

Table 4Performance comparisons (for KNN we set k to 30)

Datasets	GDA	NB	KNN	ME	LDA	SVM
20NG	95.03	85.60	50.70	89.06	93.90	91.07
WebKB4	94.01	85.13	37.29	91.93	90.72	92.04
WebKB	79.02	61.01	44.81	71.30	77.35	78.89
Industry sector	66.57	56.32	39.48	58.84	66.49	65.96
Reuters-Top10	81.98	83.33	74.07	81.65	71.46	81.13
Reuters-2	89.82	87.88	73.22	88.56	88.65	92.43
CSTR	88.50	90.85	82.53	92.39	68.29	88.71
K-dataset	88.44	86.14	58.26	86.19	77.69	91.90
TDT2	90.54	91.59	86.63	89.18	88.41	93.85

**Fig. 5.** Performance comparison.

involving SVM we used SVMtorch (Collobert & Bengio, 2001),⁵ which uses the one-versus-the-rest decomposition. The experimental comparisons are performed on the same datasets with exactly same testing and training settings.

Table 4 and Fig. 5 show performance comparisons. GDA outperformed all the other five methods on 20NG, WebKB4, WebKB and Industry Sector. SVM performed the best on Reuters-2, K-dataset, and TDT2. GDA outperformed LDA on all the experiments, and the improvement was significant (more than 10%) when the sample size was relatively small (in the case of CSTR, Reuters-Top10, and K-dataset).

On 20NG, the performance of GDA is 95.03%, which is approximately 10% higher than that of NB, 6% higher than that of ME, and 4% higher than that of SVM. On the WebKB4 dataset, GDA beats NB by approximately 5%, and both ME and SVM by approximately 2%. On the WebKB dataset, GDA beats NB by approximately 16% and ME by 6%. The performance of GDA is about 8% higher than that of NB and by 6% than that of ME on the Industry Sector. The results with GDA and with SVM are almost the same on WebKB, Industry Sector, Reuters-Top10, and CSTR. On Reuters-2, K-dataset, and TDT2, SVM performs slightly better than GDA by 3%. ME achieves the best results on the CSTR dataset while NB is the winner on Reuters-top10 in terms of performance. On CSTR, the performance of GDA is 2% lower than that of NB and 4% lower than that of ME. On Reuters-Top10, GDA is beaten by NB by approximately 1%. In total, the performance of GDA is always either the winner or very close to the winner: it is ranked the first four times, ranked the second three times, and ranked the third one time, and ranked the fourth one time. Although there is no single winner over all datasets, GDA seems to outperform the rest on most counts. We can say that GDA is a viable, competitive algorithm in text categorization.

GDA is also very efficient and most experiments are done in several seconds. Table 5 summarizes the running time for all the experiments of GDA and SVM. Figs. 6 and 7 present the comparisons of training and prediction time, respectively. The time saving of GDA is very obvious. There are several reasons contribute to the efficiency of GDA: First, GDA is performed on H_w and H_b which are smaller in size than the original scatter matrices S_w and S_b . Second, we only need to compute the first k columns of X in GSVD since the rank(H_b) is less than k . Including more columns in the transformation matrix G will not

⁵ Download-able at <http://old-www.idiap.ch/learning/SVMtorch.html>.

Table 5
Time in seconds

Datasets	GDA		SVM	
	Training	Prediction	Training	Prediction
20NG	171.80	6.86	270.20	64.28
WebKB4	63.4	0.20	114.67	54.72
WebKB	94.64	0.43	1108.17	103.03
Industry sector	88.23	6.45	423.54	79.82
Reuters-Top10	61.23	0.15	94.28	18.65
Reuters-2	96.19	1.13	566.53	85.10
CSTR	3.65	0.02	7.50	2.77
K-dataset	62.88	0.18	84.56	47.70
TDT2	21.69	5.14	89.91	26.76

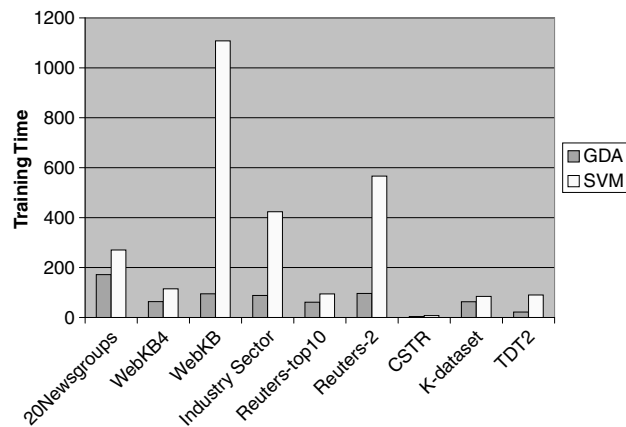


Fig. 6. Training time comparisons.

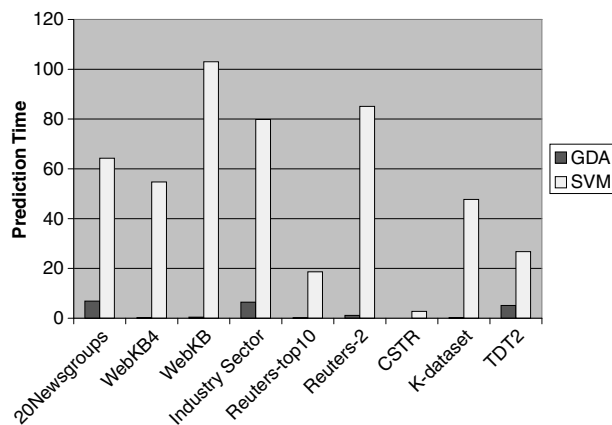


Fig. 7. Prediction time comparisons.

change the cluster structure too much; Third, in the computation of GSVD, we do not need to perform matrix inversion (see p. 472 of Golub & Loan (1996)). Finally, our approach makes use of effective feature selection via information gain, with which we can remove up to 90% or more of the unique terms without significant performance degrade (Yang & Pederson, 1997).

In summary, these experiments have shown that GDA provides an alternate choice for fast and efficient text categorization.

8. Discussions and conclusions

In this paper, we presented GDA, a simple, efficient, and yet accurate, direct approach to multi-class text categorization. GDA utilizes GSVD to transform the original data into a new space, which could reflect the inherent similarities between classes based on a new optimization criterion. Extensive experiments clearly demonstrate its efficiency and effectiveness.

Interestingly enough, although traditional discriminant approaches have been successfully applied in pattern recognition, little work has been reported on document analysis. As we mentioned earlier, this is partly because the intra-class covariance matrix is usually singular for document-term data and hence restrict the usage of discriminant. Our new criterion avoids the problem while still preserving the discriminative power of the covariance matrix. Another big barrier to application of discriminant analysis in document classification is its large computation cost. As we know, traditional discriminant analysis requires a large amount of computation on matrix inversion, SVD, and eigenvalue analysis. The costs of these operations are extremely large in document analysis because the matrices have thousands of dimension. Our approach makes use of effective feature selection via information gain, with which we can remove up to 90% or more of the unique terms without significant performance degrade (Yang & Pederson, 1997). One of our future plans is to explore how the performance correlates with different feature selection methods and the number of words selected. There are also other possible extensions such as using random projection to reduce the dimensionality before applying discriminant analysis (Papadimitriou, Tamaki, Raghavan, & Vempala, 1998).

Acknowledgements

This work is supported in part by NSF Grants EIA-0080124, DUE-9980943, and EIA-0205061, and NIH Grant P30-AG18254.

References

- Allwein, E. L., Schapire, R. E., & Singer, Y. (2000). Reducing multiclass to binary: A unifying approach for margin classifiers. In *Proceedings of the 17th international conference on machine learning (ICML 2000)* (pp. 9–16). San Francisco, CA: Morgan Kaufmann.
- Apte, C., Damerau, F., & Weiss, S. (1998). Text mining with decision rules and decision trees. In *Proceedings of the workshop with conference on automated learning and discovery: Learning from text and the web*.
- Bai, Z. (1992). The CSD, GSVD, their applications and computations. Technical report IMA preprint series 958, Minneapolis, MN.
- Chakrabarti, S., Roy, S., & Soundalgekar, M. V. (2002). Fast and accurate text classification via multiple linear discriminant projections. In *Proceedings of the 28th international conference on very large data bases (VLDB'02)*. San Francisco, CA: Morgan Kaufmann.
- Cohen, W. W., & Singer, Y. (1996). Context-sensitive learning methods for text categorization. In *SIGIR-96*.
- Collobert, R., & Bengio, S. (2001). SVM-Torch: Support vector machines for large-scale regression problems. *Journal of Machine Learning Research*, 1, 143–160.
- Deerwester, S. C., Dumais, S. T., Landauer, T. K., Furnas, G. W., & Harshman, R. A. (1990). Indexing by latent semantic analysis. *Journal of the American Society of Information Science*, 41(6), 391–407.
- Demmel, J., & Veselić, K. (1992). Jacobi's method is more accurate than QR. *SIAM Journal of Matrix Analysis and Applications*, 13, 1204–1245.
- Dieterich, T. G., & Bakiri, G. (1995). Solving multiclass learning problems via error-correcting output codes. *Journal of Artificial Intelligence Research*, 2, 263–286.
- Dumais, S., Platt, J., Heckerman, D., & Sahami, M. (1998). Inductive learning algorithms and representations for text categorization. In *CIKM*.
- Fisher, R. (1936). The use of multiple measurements in taxonomic problems. *Annals of Eugenics*, 7, 179–188.
- Fragoudis, D., Meretakakis, D., & Likothanassis, S. (2002). Integrating feature and instance selection for text classification. In *Proceedings of the 8th ACM SIGKDD international conference on knowledge discovery and data mining (SIGKDD 2002)*.
- Fukunaga, K. (1990). *Introduction to statistical pattern recognition* (2nd ed.). New York: Academic Press.
- Ghani, R. (2000). Using error-correcting codes for text classification. In *ICML-00*.
- Godbole, S., Sarawagi, S., & Chakrabarti, S. (2002). Scaling multi-class support vector machine using inter-class confusion. In *Proceedings of the 8th ACM SIGKDD international conference on knowledge discovery and data mining (SIGKDD 2002)* (pp. 513–518). New Orleans: ACM Press.
- Golub, G. H., & Loan, C. F. V. (1996). *Matrix computations*. Baltimore, MD: Johns Hopkins University Press.
- Han, E.-H., Boley, D., Gini, M., Gross, R., Hastings, K., Karypis, G., et al. (1998). WebACE: A web agent for document categorization and exploration. In *Proceedings of the 2nd international conference on autonomous agents (Agents'98)*. New York, USA: ACM Press.
- Hastie, T., & Tibshirani, R. (1998). Classification by pairwise coupling. In M. I. Jordan, M. J. Kearns, & S. A. Solla (Eds.), *Advances in neural information processing systems* (Vol. 1, pp. 507–513). Cambridge, MA: MIT Press.
- Howland, P., & Park, H. (2003). Generalizing discriminant analysis using the generalized singular value decomposition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*.
- Joachims, T. (1998). Making large-scale support vector machine learning practical. In *Advances in kernel methods: Support vector machines*.
- Joachims, T. (2001). A statistical learning model of text classification with support vector machines. In W. B. Croft, D. J. Harper, D. H. Kraft & J. Zobel (Eds.), *Proceedings of the 24th ACM international conference on research and development in information retrieval (SIGIR'01)* (pp. 128–136). New York, USA: ACM Press.
- Lam, W., & Ho, C. (1998). Using a generalized instance set for automatic text categorization. In *SIGIR-98* (pp. 81–89).
- Lewis, D. D. (1998). Naive (Bayes) at forty: The independence assumption in information retrieval. In *ECML-98*.
- Li, T., Zhu, S., & Ogihara, M. (2003). Efficient multi-way text categorization via generalized discriminant analysis. In *Proceedings of 12th international conference on information and knowledge management (CIKM 2003)* (pp. 317–324). New York, USA: ACM Press.
- Loan, C. F. V. (1976). Generalizing the singular value decomposition. *SIAM Journal of Numerical Analysis*, 13, 76–83.
- Masand, B., Linoff, G., & Waltz, D. (1992). Classifying news stories using memory based reasoning. In *SIGIR-92* (pp. 59–64).
- McCallum, A. K. (1996). Bow: A toolkit for statistical language modeling, text retrieval, classification and clustering. <<http://www.cs.cmu.edu/mccallum/bow/>>.
- Ng, H. T., Goh, W. B., & Low, K. L. (1997). Feature selection, perceptron learning, and a usability case study for text categorization. In *SIGIR-97*.
- Nigam, K., Lafferty, J., & McCallum, A. (1999). Using maximum entropy for text classification. In *IJCAI-99 workshop on machine learning for information filtering*.
- Papadimitriou, C. H., Tamaki, H., Raghavan, P., & Vempala, S. (1998). Latent semantic indexing: A probabilistic analysis. In *Proceedings of the 17th ACM SIGACT-SIGMOD-SIGART symposium on principles of database systems* (pp. 159–168).
- Porter, M. F. (1997). *An algorithm for suffix stripping*, 313–316.

- Schapire, R. E., & Singer, Y. (2000). Boostexter: A boosting-based system for text categorization. *Machine Learning*, 39(2/3), 135–168.
- Scholkopf, B., & Smola, A. J. (2002). *Learning with kernels*. Cambridge, MA: MIT Press.
- TD2 (1998). Nist topic detection and tracking corpus. <<http://www.nist.gov/speech/tests/tdt/tdt98/index.htm>>.
- Tzeras, K., & Hartmann, S. (1993). Automatic indexing based on Bayesian inference networks. In *SIGIR-93*.
- Vapnik, V. N. (1998). *Statistical learning theory*. New York: Wiley.
- Wiener, E. D., Pedersen, J. O., & Weigend, A. S. (1995). A neural network approach to topic spotting. In *Proceedings of the 4th annual symposium on document analysis and information retrieval*.
- Yang, Y., & Liu, X. (1999). A re-examination of text categorization methods. In *Proceedings of the 22nd annual international conference on research and development in information retrieval (SIGIR'99)* (pp. 42–49). New Orleans: ACM Press.
- Yang, Y., & Pederson, J. O. (1997). A comparative study on feature selection in text categorization. In *Proceedings of the 14th international conference on machine learning (ICML-97)* (pp. 412–420). San Francisco, CA: Morgan Kaufmann.