Optimally Combining Positive and Negative Features for Text Categorization

Zhaohui Zheng ZZHENG3@CEDAR.BUFFALO.EDU
Rohini Srihari ROHINI@CEDAR.BUFFALO.EDU
CEDAR, Dept. of Computer Science and Engineering, State University of New York at Buffalo, NY 14260 USA

Abstract

This paper presents a novel local feature selection approach for text categorization. It constructs a feature set for each category by first selecting a set of terms highly indicative of membership as well as another set of terms highly indicative of non-membership, then unifying the two sets. The size ratio of the two sets was empirically chosen to obtain optimal performance. This is in contrast with the standard local feature selection approaches that either (1) only select the terms most indicative of membership; or (2) implicitly but not optimally combine the terms most indicative of membership with non-membership. The experimental comparison between the proposed approach and standard approaches was conducted on four feature selection metrics: chi-square, correlation coefficient, odds ratio, and GSS coefficient. The results show that the proposed approach improves text categorization performance.

1. Introduction

Text categorization is a machine learning task, defined as automatically assigning predefined category labels to new free text documents. A growing number of statistical machine learning techniques have been applied to text categorization in recent years, notable among which are five approaches: nearest neighbor classifier (Creey et al., 1992; Yang, 1994), Bayesian classifier (Tzeras & Hartman, 1993; Lewis & Ringuette, 1994), decision tree (Apte, Damerau, & Weiss, 1994), neural networks (wiener, Pederson & Weigend, 1995; Ng, Goh & Low, 1997), and support vector machines (Joachims, 1998).

One major difficulty in text categorization problems is the high dimensionality of the input feature space typical for textual data. This is because each distinct term or token appearing in the document collection represents one dimension in the feature space. For a typical document collection, there are tens of thousands or even hundreds of thousands of distinct terms or tokens. After the elimination of stop words and stemming, the set of features is still too large for many learning algorithms, e.g. neural networks. In order to improve scalability of text categorization, we need to apply feature selection techniques to reduce the feature size further more. Various feature selection methods have been proposed in the literature and their relative merits have been tested by experimentally evaluating the text categorization performance. There are two distinct ways of viewing feature selection, depending on whether the task is performed locally or globally: (1) local feature selection. For each category, a set of terms is chosen for classification based on the relevant and irrelevant documents in this category. (2) global feature selection. A set of terms is chosen for the classification under all categories based on the relevant documents in the categories. The local feature selection for each category can be viewed as the global feature selection for two "categories": *relevant* and *irrelevant*. Local feature selection is of interest in this paper.

Several feature selection measures have been explored in the literature including Document Frequency (DF), Information Gain (IG), Mutual Information (MI), Chi-square (CHI), Correlation Coefficient (CC), Odds ratio (OR) and GSS coefficient (GSS) (Galavotti, Sebastiani, & Simi, 2000; Mitchell, 1996; Mladeni, 1998; Ng, Goh & Low, 1997; Quinlan, 1986; Rijsbergen, 1979; Sebastiani, 2002; Schutze, Hull & Pederson, 1995; Yang & Pedersen, 1997). Out of the seven measures, CHI, CC, OR and GSS seem to be the most effective based on the experiments reported so far. We will focus on the four measures.

This paper presents a novel local feature selection method that explicitly selects and combines the features highly indicative of membership and non-membership for each category in a way such that the optimal performance, e.g. F1 measure, will be obtained on a validation set. The features indicative of membership and non-membership are also referred to as the positive and negative features respectively. The presence of positive and negative features in a document indicates its relevance and non-relevance respectively.

The rest of the paper is organized as follows. Section 2

Workshop on Learning from Imbalanced Datasets II, ICML, Washington DC, 2003.

describes the four feature selection measures and standard methods of using them. Section 3 presents the proposed feature selection technique. In Section 4, we describe naïve Bayes classifier whose performance will be used to evaluate the effectiveness of various feature selection methods. Experimental results are analyzed in Section 5. Conclusions are given in Section 6.

2. Related Work

In this section, we will first briefly review the four feature selection measures, then present the methods of using them in the literature, and finally describe the imbalanced data problem and its impacts on feature selection. Note that, the methods here refer to the scheme of applying feature selection measures to term selection.

2.1 Feature Selection Measures

In what follows, A, B, C, and D will denote the numbers of times a term *t* and a category c_i co-occur, *t* occurs without c_i, c_i occurs without *t*, and neither c_i nor *t* occurs, respectively. N represents the total number of documents.

2.1.1 Chi-Square (CHI)

CHI measures the lack of independence between a term *t* and a category c_i and can be compared to the chi-square distribution with one degree of freedom to judge extremeness (Yang, 1999; Schutze, Hull & Pederson, 1995). It is defined as:

$$\chi^2(t,c_i)=\frac{N[P(t,c_i)\cdot P(\bar{t},\bar{c}_i)-P(t,\bar{c}_i)\cdot P(\bar{t},c_i)]^2}{P(t)\cdot P(\bar{t})\cdot P(c_i)\cdot P(\bar{c}_i)}$$

$$\approx\frac{N(AD-CB)^2}{(A+C)\times(B+D)\times(A+B)\times(C+D)}$$

χ^2 has a natural value of zero if *t* and c_i are independent. It is a normalized value, and hence is comparable across terms for the same category.

2.1.2 Correlation Coefficient (CC)

Correlation coefficient $CC(t,c_i)$ of a word *t* with a category c_i was defined by Ng et al. as (Ng, Goh & Low, 1997; Sebastiani, 2002):

$$CC(t,c_i)=\frac{\sqrt{N}[P(t,c_i)\cdot P(\bar{t},\bar{c}_i)-P(t,\bar{c}_i)\cdot P(\bar{t},c_i)]}{\sqrt{P(t)\cdot P(\bar{t})\cdot P(c_i)\cdot P(\bar{c}_i)}}$$

$$\approx\frac{\sqrt{N}(AD-CB)}{\sqrt{(A+C)\times(B+D)\times(A+B)\times(C+D)}}$$

It is a variant of the CHI metric, where $CC^2=\chi^2$. CC can be viewed as a "one-sided" chi-square metric. The positive values correspond to features indicative of membership, while negative values indicate non-membership. The greater (smaller) the positive (negative) values are, the stronger the terms will be to indicate the membership (non-membership). Standard CC based local feature selection method selects the terms with maximum CC value as features. The rationale behind is that terms coming from the irrelevant texts of a category are considered useless. On the other hand, CHI is non-negative, whose values indicate the membership or non-memberships of a term to one category. Accordingly the ambiguous features will be ranked lower. In contrast with CC, CHI considers the terms coming from both the relevant and non-relevant texts.

2.1.3 Odds Ratio (OR)

Odds ratio was proposed originally by van Rijsbergen et al. (1979) for selecting terms for relevance feedback. The basic idea is that the distribution of features on the relevant documents is different from the distribution of features on the non-relevant documents. It has been used by Mladenic (1998) for selecting terms in text categorization. It is defined as follows:

$$OR(t,c_i)=\frac{P(t|c_i)\cdot[1-P(t|\bar{c}_i)]}{[1-P(t|c_i)]\cdot P(t|\bar{c}_i)}\approx\frac{AD}{CB}$$

The values greater than 1 correspond to features indicative of membership, while the values less than 1 correspond to features indicative of non-membership. It only considers the terms from relevant text. The Expected Likelihood Estimate (ELE) smoothing technique was used in this paper to handle singularities:

$$OR(t,c_i)\approx\frac{\frac{A+0.5}{A+C+1}\cdot(1-\frac{B+0.5}{B+D+1})}{(1-\frac{A+0.5}{A+C+1})\cdot\frac{B+0.5}{B+D+1}}=\frac{(A+0.5)\cdot(D+0.5)}{(C+0.5)\cdot(B+0.5)}$$

2.1.4 GSS Coefficient

GSS Coefficient is another simplified variant of the χ^2 statistics proposed by Galavotti et al. (2000), which is defined as:

$$GSS(t,c_i)=P(t,c_i)\cdot P(\bar{t},\bar{c}_i)-P(t,\bar{c}_i)\cdot P(\bar{t},c_i)$$

$$\approx\frac{AD-CB}{N^2}$$

Similar to CC, the positive values correspond to features indicative of membership, while negative values indicate non-membership. Therefore, only the positive terms are considered.

2.2 Feature Selection Methods

Each of the above four measures is actually a function $f(t, c_i)$ with a term t and a category c_i as its parameters. The value indicates some relationship between the term and the category. In global feature selection, we assess the value of a term in a global, or category-independent, sense. Either the average or the maximum of their category-specific values are usually computed and compared (Yang & Pederson, 1997). That is,

$$f_{avg}(t) = \sum_{i=1}^{m} P(c_i) f(t, c_i)$$

$$f_{\max}(t) = \max_{i=1}^{m} \{ f(t, c_i) \}$$

Given a vocabulary V and a function f that maps terms to real values, we define two subsets of V of size l, viz., $Max[V, f, l] \subseteq V$ and $Min[V, f, l] \subseteq V$ as follows:

$$\forall x \in Min[V, f, l], \forall y \in V - Min[V, f, l], \quad f(x) \le f(y),$$

$$\forall x \in Max[V, f, l], \forall y \in V - Max[V, f, l], \quad f(x) \ge f(y)$$

In other words, $Max[V, f, l]$ and $Min[V, f, l]$ consists of the l terms $t_j \in V$ with the highest and lowest $f(t_j)$ values respectively. Then, in global feature selection, the feature set F will be $Max[V, f_{\max}, l]$ or $Max[V, f_{avg}, l]$, where l is the size of F.

In local feature selection, a feature set is constructed for each category. Accordingly, the feature set F_i for c_i will be $Max[V, f(\ , c_i),\ l]$, where f can be any feature selection measure that uses two way contingency table of a term *t* and a category c_i.

Local feature selection methods using asymmetric measures, e.g. CC, OR and GSS, actually pick out the terms most indicative of membership. They will never consider negative features unless all the positive features have already been selected.

On the other hand, local feature selection methods using symmetric measures, e.g. CHI, implicitly combine the terms most indicative of membership and non-membership. The size ratio between the positive and negative features is internally decided by thresholding on the size of feature set.

2.3 Imbalanced Data Problem

When training a binary text classifier (text filtering system) for a category, we use all the documents in the training corpus that belong to that category as relevant training data and all the documents in the training corpus that belong to all the other categories as non-relevant training data. It is often the case that there is an overwhelming number of non-relevant training documents especially when there is a large collection of categories with each assigned to a small number of documents. Many approaches have been employed to address the imbalanced data problem. The concepts of "query zone" and "category zone" were introduced to select a subset of the non-relevant documents as the non-relevant training data (Hearst, et al., 1996; Ruiz & Srinivasan, 1999). These documents are the most relevant non-relevant documents. Essentially, these methods try to obtain more balanced relevant and non-relevant training documents. In this paper, we consider this problem from a different perspective. Instead of balancing the training data, our method balances the positive and negative features, e.g., generates the optimal combination of positive and negative features according to the imbalanced data.

The impacts of imbalanced data problem on the standard local feature selection for text filtering can be illustrated as follows:

(1) For the methods using the positive features only (e.g. CC, OR, or GSS), the non-relevant documents are subject to misclassification. It will be even worse for the imbalanced data problem, where non-relevant documents dominate. How to confidently reject the non-relevant documents is very important in that case.

(2) When applying to the imbalanced data the methods implicitly combining the positive and negative features, e.g. CHI, the positive features usually have much higher values than the negative features according to its definition and our previous experiments. Therefore, the positive features will dominate in the feature set. The similar situation occurs as described in (1). For example, the upper limit CHI value of a positive or negative feature is $\frac{N \cdot (A+C) \cdot (B+D)}{(A+B) \cdot (C+D)}$. For the positive feature, it represents the case that the feature appears in every relevant document, but never in any non-relevant document. For the negative feature, it means that the feature appears in every non-relevant document, but never in any relevant document. Due to the large amount and diversity of the non-relevant documents in imbalanced data set, it is much more difficult for a negative feature to achieve that

maximum than a positive feature. This extreme example shed light on why the CHI values of positive features are usually much larger than that of negative features. It will be inappropriate of standard local feature selection using CHI to simply compare their CHI values without considering whether they are positive or negative.

3. Combining Positive and Negative Features

We believe that:

(1) The negative features are also useful and should be included in the feature set. Since the local feature selection can be viewed globally by considering relevant and non-relevant as two "categories", the negative features are actually from the "non-relevant" category. In such a bi-category problem, intuitively thinking, terms from both of them should be considered. The presence of negative features in a document is a good indicator of its non-membership. Thus the text filtering performance can be improved through confident rejection of non-relevant documents.

(2) Implicit combination of positive and negative features is not necessarily optimal especially for imbalanced data set, in which the values of positive features are usually much larger than negative features. CHI might only select the positive features (equivalent to standard CC based approach in this case) when the size of feature set is small. Thus the size ratio of the positive and negative features should be explicitly set and empirically tuned to different scenario: data collection, text classifier, etc.

Based on the above two observations, we propose a new feature selection approach containing the following three steps:

For each category c_i:

Step 1: generate a positive-feature set F_i^+ as $Max[V, f(\cdot, c_i), l_1]$, $l_1, 0 < l_1 \le l$, is a nature number.

Step 2: generate a negative-feature set F_i^- as $Max[V, f(\cdot, \overline{c}_i), l_2]$, $l_2 = l - l_1$ is a non-negative integer.

Step 3: $F_i = F_i^+ \cup F_i^-$.

Where: l, $l << |V|$, is the predefined size of feature set. l_1/l, $0 < l_1/l \le 1$, is the key parameter and should be chosen to optimize the categorization performance on a validation set. When $l_1 = l$, e.g. $l_2 = 0$, the method corresponds to the standard local feature selection method. So, the standard method can be viewed as one particular case of our method.

In Step 1, we intend to pick out those terms most indicative of membership of c_i, while in step 2, those terms most indicative of non-membership are selected as well. The feature set will be the union of the two.

Accordingly, the function $f(t, c_i)$ should satisfy: the larger the function value is, the more likely the term belongs to the category c_i. Obviously, CC, OR, and GSS can serve as such functions, while CHI can not. The reasons why we present CHI in this paper are as follows:

(1) CHI has been proved to be an effective and robust feature selection measure in the literature. In order to make our experiments comparable to others, we use it as our baseline.

(2) CHI is very related in concept to CC based approaches using either the standard method or our proposed method, as will be shown later.

Based on the definition of the three measures, we can easily obtain:

$$CC(t, \overline{c}_i) = -CC(t, c_i),$$
$$OR(t, \overline{c}_i) = \frac{1}{OR(t, c_i)},$$
$$GSS(t, \overline{c}_i) = -GSS(t, c_i).$$

Accordingly, Step 2 can be rewritten as:

Step 2: generate a negative-feature set F_i^- as $Min[V, f(\ , c_i), l_2]$.

Compared with the standard methods that only consider the terms indicative of membership, e.g., CC, OR and GSS, we add the step 2, which add to the feature set those terms indicative of non-membership. The advantage of our approach over the standard one can be illustrated by the following simple example: given a list of terms *t1*, *t2*,…, *t8* whose CC values are 9, 8.5, 8.2, 8, 1, -1, -5.8, and -5.9 respectively. If the size of feature set is 6, *t1* through *t6* will be selected. Suppose a new document containing *t5*, *t7* and *t8* comes in; the system will assign it as relevant although it is irrelevant. On the other hand, the proposed approach will be more likely to choose *t7* and *t8* instead of *t5* and *t6* and hence classify the new document correctly.

When applying our method to CC, the resulted approach seems very similar to the standard CHI based approach:

(1) Both of them consider not only the terms indicative of membership but non-membership also. The proposed method using CC explicitly combines them while standard CHI implicitly considers them.

(2) Because CHI value is equal to the squared CC value, among those terms with positive/negative CC value indicative of membership/non-membership, the greater/smaller the value is, the more likely it will be selected as features by both methods.

However, the major differences between two approaches are:

(1) CHI does not differentiate between the terms indicative of membership and non-membership by comparing the squared values. Although it might consider in concept both positive and negative features, the size ratio between them is not optimal. There are no extra parameters to optimize that ratio. In contrast, due to its design, our approach can optimize the size ratio to get best performance. Let us refer to the above example. If we apply CHI to select four features, *t1* through *t4* will be selected, each of which is from relevant document set. When the same new document comes in, the system can hardly tell whether it is relevant or not.

(2) Because the positive examples are far fewer than the negative examples in the training corpus, CHI actually favors the positive features according to its definition. In other words, the CHI values are not comparable between the positive and the negative features. Usually the values of positive features are much larger than negative features as described in Section 2.3. The proposed approach, however, allows the sizes of the feature set to be as small as needed while guaranteeing that the system uses both positive and negative features in an optimal way.

4. Naïve Bayes Classifier for Text Filtering

Naïve Bayes classifier is a highly practical Bayesian learning method [6]. The central idea is to use the joint probabilities of terms and categories to estimate the probabilities of categories given a document. The naïve part of such a model is the simplifying assumption that the words are conditionally independent given the category as well as the probability of word occurrence is independent of position within the text. For text filtering, the relevance score between a new document d and the category c can be calculated as:

$$Score(d,c) = \frac{\log P(c) + \sum_{\omega_i} \log P(\omega_i \mid c)}{\log P(\bar{c}) + \sum_{\omega_i} \log P(\omega_i \mid \bar{c})}$$

where: ω_i is the feature appearing in the document d ;

$P(c)$ and $P(\bar{c})$ represent the prior probabilities of relevant and non-relevant respectively;

$P(\omega_i \mid c)$ and $P(\omega_i \mid \bar{c})$ represent the likelihood probabilities of ω_i appearing in relevant and non-relevant training documents respectively.

A binary decision (relevant or non-relevant) on d with respect to the category c is obtained by thresholding on $Score(d,c)$. We train one naïve Bayes classifier per category. A relevance score threshold is learned per category to empirically optimize F1 measure on the validation set.

5. Experimental Results and Analysis

5.1 Experimental Setting

To make our results comparable to others, we have used the Reuters-21578 corpus (Yang, 1999; Yang & Pederson, 1997), as it is a widely used benchmark in text categorization domain. For this paper, we use the ApteMod version of Reuters-21578 as described by Yang (1999). Finally we obtain 90 categories in both the training and test sets, a training set of 7,769 documents, and a test set of 3,019 documents. The average number of categories per document is 1.3. The number of positive instances per category ranges from a minimum of 1 to maximum of 2,877 in the training set. In order to automatically learn the category specific parameters, e.g. size ratio in feature selection and thresholds in classification, we use two thirds of the training set for training and the remaining one third as "validation". After obtaining these thresholds, the classifiers will be retrained on the whole training set.

Classification effectiveness has been evaluated in terms of the standard precision, recall and F1 measure.

The precision, recall and F1 for each category c_i are defined as:

$$R_i = \frac{\alpha_i}{\beta_i}, \quad P_i = \frac{\alpha_i}{\gamma_i}, \quad F1_i = \frac{2P_iR_i}{P_i + R_i},$$

where: α_i is the number of documents correctly assigned by system to category c_i , and

β_i is the number of documents assigned by system to category c_i , and

γ_i is the number of documents from category $c_i \quad (i = 1,2,\cdots m)$.

These category-relative values may in turn be averaged according to two alternative ways:

(1) macro-averaging: the precisions and recalls can be computed for the binary decisions on each individual category first and then be averaged over categories. That is,

$$macroP = \frac{\sum_{i=1}^{m} P_i}{m}, \quad macroR = \frac{\sum_{i=1}^{m} R_i}{m}, \text{ and}$$

$$macroF1 = \frac{\sum_{i=1}^{m} F1_i}{m},$$

(2) micro-averaging: the precisions and recalls are computed globally over all the *n* x *m* binary decisions where *n* is the number of total test documents, and *m* is the number of categories. That is,

$$microR = \frac{\sum_{i=1}^{m} \alpha_i}{\sum_{i=1}^{m} \beta_i}, \quad microP = \frac{\sum_{i=1}^{m} \alpha_i}{\sum_{i=1}^{m} \gamma_i}, \text{ and}$$

$$microF1 = \frac{2 \times microR \times microP}{microR + microP}$$

micro-averaging F1 has been widely used in cross-method comparisons. In this paper, we will focus on this measure. Accordingly, the size ratio between the positive and negative feature sets will be optimized to get best micro-averaged F1 measure on the validation set.

In order to compare our proposed feature selection approach with the standard one, we apply them to naïve Bayes classifiers. Three groups of feature selection methods are considered:

Group 1: Standard CHI, Standard CC, and improved CC. The three methods are referred as G11, G12, and G13 respectively for short form.

Group 2: Standard OR and improved OR, referred as G21 and G22.

Group 3: Standard GSS coefficient and improved GSS coefficient, referred as G31 and G32.

where: standard CHI, CC, OR and GSS represents the standard local feature selection methods using CHI, CC, OR and GSS measures respectively, while the improved CC, OR and GSS are the application of the proposed feature selection method to CC, OR and GSS measures respectively. Note that, there is no "improved CHI" method because CHI measure does not satisfy the requirement as mentioned in Section 3. However, due to its similarity with CC, we put standard CHI in the group of standard CC and improved CC. The feature selection methods are compared with each other in the same group. Typical size of a local feature set is between 10 and 50 (Sebastiani 2002). In this paper the performances are reported at the range of 10 ~ 100.

5.2 Experimental Results

Table 1 lists the micro-averaged F1 values for naïve Bayes classifiers with the seven different feature selection methods (as listed in the first row) at different sizes of feature set (as listed in the first column).

$\lvert F_i \rvert$	G11	G12	G13	G21	G22	G31	G32
10	.771	.76	.781	.628	.641	.733	.774
20	.767	.763	.803	.644	.661	.74	.78
30	.782	.765	.816	.654	.671	.74	.797
40	.778	.76	.812	.669	.687	.74	.797
50	.784	.769	.82	.689	.712	.734	.797
100	.779	.751	.819	.721	.762	.734	.802

Table 1: Micro-averaged F1 values for naïve Bayes classifiers with the seven feature selection methods at different sizes of features.

As is shown in Table 1, the improved Correlation Coefficient method (G13) is much better than the standard CC (G12) and CHI (G11) method, and the improved Odds ratio (G22) and GSS Coefficient methods (G32) greatly outperform the corresponding standard methods (G21 and G31 respectively). This confirms our intuition that by optimally combining positive features with negative features, the text categorization performance will be remarkably improved.

Table 2 lists the micro-averaged precision, recall of each method when the micro-averaged F1 is maximum over the different sizes of features. For example, G11 achieve its maximum micro-averaged F1 (.784) as the size of feature set is 50 according to the first two columns of Table 1. The second row in Table 2 gave the corresponding micro-averaged precision and recall as well. From Table 2, we can see our proposed approach greatly increases the micro-averaged recall and F1 without hurting precision too much. Because we optimize F1 measure for each category, the more balanced micro-averaged precision and recall are obtained. It also explains why the micro-averaged precision remains unimproved.

In order to illustrate the ratio of negative features in the feature set, we list in Table 3 the number of categories, in which the number of positive features is greater than, smaller than or equal to the number of negative features in case of improved CC (feature size = 50). The three cases

Method	microP	MicroR	microF1
G11	.843	.732	.784
G12	.84	.709	.769
G13	.818	.822	.82
G21	.744	.70	.721
G22	.734	.792	.762
G31	.793	.695	.74
G32	.786	.818	.802

Table 2: Micro-averaged precision, recall and F1 values for naïve Bayes classifiers with the seven feature selection methods.

correspond to l_1 / l > 0.5, < 0.5 and = 0.5 respectively in the first column of Table 3. Table 3 shows that in order to obtain best text categorization performance in terms of F1, we should select more negative features than positive features in 47 out of the 90 categories. It reconfirms the usefulness of negative features. Our explanation is: when the negative examples are overwhelming, rejection of the negative examples with high confidence (accuracy) will be of more importance, which could be achieved by increasing the number of the negative features.

l_1 / l	Number of categories
> 0.5	33
< 0.5	47
= 0.5	10

Table 3: The number of categories in which the size of positive set is greater than, smaller than or equal to the negative set in the case that the improved CC obtain best performance (feature size is 50.)

6. Conclusions

Experiments with four known feature selection measures and methods and a new feature selection method have been described. We proposed an effective feature selection method that optimally combines the terms most indicative of membership and non-membership. The main conclusions are:

- The terms indicative of non-membership are useful and should be considered in local feature selection.
- By explicitly and optimally setting the size ratio of the positive and negative features, the text categorization performance was improved greatly.

References

C. Apte, F. Damerau, & S. Weiss (1994). Towards language independent automated learning of text categorization models. In *Proceeding of the 17th Annual ACM/SIGIR conference*.

R.H. Creecy, et al. (1992). Trading mips and memeory for knowledge engineering: classifying census returns on the connection machine. Comm. ACM, 35:48-63.

Galavotti, L., Sebastiani, F., & Simi, M. (2000). Experiments on the use of feature selection and negative evidence in automated text categorization. In *Proceedings of ECDL-00, 4th European Conference on Research and Advanced Technology for Digital Libraries* (Lisbon, Portugal, 2000), 59-68.

Marti Hearst, et al. (1996). Xerox TREC4 site report. In *Proceedings of the Fourth Text Retrieval Conference TREC-4*.

Thorsten Joachims (1998). Text categorization with Support Vector Machines: Learning with Many Relevant Features. In *European Conference on Machine Learning (ECML),* pages 137-142, Berlin, Springer.

D.D. Lewis & M. Ringuette (1994). Comparison of two learning algorithms for text categorization. In *Proceedings of the Third Annual Symposium on Document Analysis and Information Retrieval (SIGIR'94).*

Tom Mitchell (1996). *Machine Learning*. McGraw Hill.

Mladeni D. [1998]. *Machine Learnimg on non-homogeneous, distributed text data*. PhD Dissertation, Univeristy of Ljubljana, Slovenia, 1998.

H.T. Ng, W.B. Goh, & K.L. Low (1997). Feature selection, perceptron learning, and a usability case study for text categorization. In *20th Ann Int ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR'97),* pages 67-73.

J.R. Quinlan (1986). Induction of decision trees. Machine Learning, 1(1):81-106.

Van Rijsbergen CJ (1979). *Information Retrieval*. Butterworths, London, 2nd edition.

Ruiz, M.E. & Srinivasan, P. (1999). Hierarchical neural networks for text categorization. In *Proceedings of SIGIR-99, 22nd ACM International Conference on Research and Development in Information Retrieval*, 281-282.

Sebastiani F. (2002). Machine Learning in Automated Text Categorization. *ACM Computing Surveys*, Vol 34, No. 1, pp. 1-47.

Schutze H, Hull D.A. & Pederson J.O. (1995). A comparison of classifiers and document representations for the routing problem. In *Proceedings of the 18th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, Seattle, WA, pp. 229-337.

K. Tzeras & S. Hartman (1993). Automatic indexing based on bayesian inference networks. In *Proc 16th Int ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR'93)*, pages 22-34.

A.S. Weigend, E.D. Wiener, & J.O. Pederson (1999). Exploiting hierarchy in text categorization. *Inforamtion Retrieval,* 1(1-2):6990.

E. Wiener, J.O. Pederson & A.S. Weigend (1995). A neural network approach to topic spotting. *In Proceedings of the Fourth Annual Symposium on Document Analysis and Information Retrieval (SDAIR'95)*, pages 317-332, Nevada, Las Vegas. University of Nevada, Las Vegas.

Y. Yang (1994). Expert network: Effective and efficient learning from human decisions in text categorization and retrieval. In *17th Ann Int ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR'94)*, pages 13-22.

Y. Yang (1999). An evaluation of statistical approaches to text categorization. *Journal of Information Retrieval*, 1(1/2):67-88.

Y. Yang and J.P. Pedersen (1997). A comparative study on feature selection in text categorization. In Jr. D. H. Fisher, editor, *The Fourteenth International Conference on Machine Learning*, pp. 412-420. Morgan Kaufmann.