Introducing a Family of Linear Measures for Feature Selection in Text Categorization

Elías F. Combarro, Elena Montañés, Irene Díaz, José Ranilla, and Ricardo Mones

Abstract—Text Categorization, which consists of automatically assigning documents to a set of categories, usually involves the management of a huge number of features. Most of them are irrelevant and others introduce noise which could mislead the classifiers. Thus, feature reduction is often performed in order to increase the efficiency and effectiveness of the classification. In this paper, we propose to select relevant features by means of a family of linear filtering measures which are simpler than the usual measures applied for this purpose. We carry out experiments over two different corpora and find that the proposed measures perform better than the existing ones.

Index Terms—Text categorization, feature selection, filtering measures, machine learning.

1 Introduction

Nowadays, with the wide availability of text documents in electronic form, it is of great importance to develop methods for the automatic processing of large collections of text files.

One of the main tasks in this processing is that of assigning the documents of a corpus to a set of previously fixed categories, what is known as Text Categorization (TC) [1]. This process involves some understanding of the contents of the documents and/or some previous knowledge of the topics. For this reason, this task has been traditionally performed by human readers. However, this approach is infeasible when a huge number of documents is involved. In the last decade, there has been a great effort in the research of algorithms that automatically label new documents into a set of prefixed categories, perhaps using knowledge acquired from a smaller set of documents already classified [1].

Thus, documents need to be represented in a way suitable for automatic processing. The most widely adopted representation is the *bag of words* [2], where each document is identified with a vector of real numbers. Each component corresponds to a word occurring in the corpus and its value measures the importance of the word in the document.

Obviously, the number of different words appearing in the collection tends to be very large, and the task of processing such document vectors can become infeasible. Also, it is known that most words are irrelevant for the classification task and some of them even introduce noise that may decrease the overall performance [2].

Hence, it is usual to reduce the number of words used to represent the documents in order to increase both the efficiency and the effectiveness of the classification [1]. In addition to elimination of *stop-words* and *stemming* [2], feature reduction is usually performed.

A common approach for feature reduction is Feature Selection (FS), which consists in choosing a subset of the original features for the document representation. According to John et al. [3], there are two main approaches to FS: filtering and wrapping. In the former, a feature subset is selected independently of the learning method that will use it. In the latter, a feature subset is selected using an evaluation function based on the same learning algorithm that will consider this subset of features. Although wrapper approaches have been shown to perform better [3], they can be rather time-consuming and it is sometimes infeasible to use them. For this reason, the use of filtering measures is prominent in TC. It is based on scoring the features with some relevance measures and selecting a predefined number from the top ranked [4].

In this paper, we introduce a new family of filtering measures for FS in TC. They are simpler than most existing measures, but the experiments carried out show that they perform equal or better than the most promising ones up to the moment (see Section 6).

The rest of the paper is organized as follows: In Section 2, some of the approaches adopted for FS in the past are summarized. The new measures are introduced in Section 3 at the same time that some of their properties are studied. Section 4 is devoted to describe the system used for TC, including the document preprocessing and the classifier chosen. In Section 5, the settings of our experiments, the document collections taken, and the way of quantifying the performance are described. In Section 6, we present and discuss our results and, finally, in Section 7, some conclusions and ideas for further research are detailed.

2 Previous Work

FS is a widely adopted approach for feature reduction in TC [1]. It is based on selecting a subset of features from the original set of words. FS can be performed by keeping the words with highest score according to a predetermined measure of word relevance, or filtering measure. In the rest of the section, the filtering measures previously applied for FS in TC are described.

• *The authors are with the Artificial Intelligence Center of the University of Oviedo, Edificio de Marina Civil, Campus de Viesques S/N, 33204, Gijón, Spain. E-mail: {elias, elena, sirene, ranilla, mones}@aic.uniovi.es.*

Manuscript received 19 Nov. 2004; revised 18 Mar. 2005; accepted 30 Mar. 2005; published online 19 July 2005.
For information on obtaining reprints of this article, please send e-mail to: tkde@computer.org, and reference IEEECS Log Number TKDESI-0420-1104.

2.1 Term Frequency, Document Frequency, and *tfidf*

The most simple filtering measures are the *term frequency* (*tf*) and the *document frequency* (*df*). They measure the relevance of a word by means of its total number of appearances and by the number of different documents in which it appears, respectively. They can be combined into *tfidf* defined by

$$tfidf = tf \log\left(\frac{N}{df}\right),$$

where N is the number of documents in the corpus. Notice that words appearing in all the documents are considered noninformative, independently of its absolute frequency and, in general, a word occurring in many documents will have *tfidf* smaller than others, with the same *tf*, but appearing in less documents.

Despite their simple appearance, these measures perform well in many situations [5].

2.2 Information Theory Measures

Measures taken from Information Theory (IT) have been widely used because it is interesting to consider the distribution of a word over the different categories. Among these measures, *information gain* (IG) takes into account the presence of the word in a category as well as its absence and can be defined by (see, for instance, [4])

$$IG(w,c) = P(w)P(c|w)\log\left(\frac{P(c|w)}{P(c)}\right) + P(\overline{w})P(c|\overline{w})\log\left(\frac{P(c|\overline{w})}{P(c)}\right),$$

where $P(w)$ is the probability that the word w appears in a document, $P(c|w)$ is the probability that a document belongs to the category c knowing that the word w appears in it, $P(\overline{w})$ is the probability that the word w does not appear in a document, and $P(c|\overline{w})$ is the probability that a document belongs to the category c if we know that the word w does not occur in it. Usually, these probabilities are estimated by means of the corresponding relative frequencies. In the same direction, *expected cross entropy for text* (CET) [6] only takes into account the presence of the word in a category. It is defined by

$$CET(w,c) = P(w)\cdot P(c|w)\cdot \log\frac{P(c|w)}{P(c)}.$$

Yang and Pedersen [4] introduced the statistic χ^2 for feature reduction, which measures the lack of independence between a word and a category. A modification of this measure is $S-\chi^2$, proposed by Galavotti et al. [7], who defined it by

$$S-\chi^2(w,c) = P(w,c)\cdot P(\overline{w},\overline{c}) - P(w,\overline{c})\cdot P(\overline{w},c).$$

It has been shown that it performs better than χ^2 (see [7]).

2.3 Machine Learning Measures

In [8], [9], we proposed several measures taken from the Machine Learning (ML) environment. Specifically, they are measures previously applied to quantify the quality of the rules induced by an ML algorithm. In order to be able to adopt these measures for FS in TC, we have proposed to associate to each pair of word w and category c, the following rule:

If the word w appears in a document, then that document belongs to category c.

We denote this rule by $w \rightarrow c$. In this way, the quantification of the importance of a word w in a category c was reduced to the quantification of the quality of the rule $w \rightarrow c$.

Some notation must be introduced in order to define this family of measures. For each pair (w,c), $a_{w,c}$ denotes the number of documents of the category c in which w appears, $b_{w,c}$ denotes the number of documents that contain the word w but do not belong to the category c, $c_{w,c}$ denotes the number of documents of the category c in which w does not appear and, finally, $d_{w,c}$ denotes the number of documents which neither belong to category c nor contain the word w.

In general, rule quality measures are based on the percentage of successes and failures of its application. A measure of this kind is the Laplace measure. Basically, this measure modifies the percentage of success and it is defined by

$$L(w \rightarrow c) = \frac{a_{w,c}+1}{a_{w,c}+b_{w,c}+s}, \tag{1}$$

where s is the number of classes in the classification problem. Since we will be dealing with binary problems, we will use $s=2$.

Another measure of this kind is the *difference* (D). It establishes a balance between documents of the category c and the rest, both containing the word w:

$$D(w \rightarrow c) = a_{w,c} - b_{w,c}. \tag{2}$$

This measure is a simplification of the accuracy, defined in [10]. It penalizes those words which appear in documents of the category c by means of subtracting from them the number of documents of the rest of categories which contain the word w.

Besides, there exist other rule quality measures that also take into account the number of documents of the category in which the word occurs and the distribution of the documents over the categories. An example of these kind of measures is the *impurity level* (IL) studied for FS in [8], [9].

With this measure, the distribution of documents over the categories is considered by introducing the concept of *canonical* or *unconditional rule* (denoted by $\rightarrow c$), which says that any document belongs to the category c. This rule is used as a reference for the rest of rules of the same category. The measure IL is calculated by taking into account the times that the rule is applied (n) and its successes (m). The confidence interval of its percentage of success is determined by

$$CI_{l,r} = \frac{p+\frac{z^2}{2n} \pm z\sqrt{\frac{p(1-p)}{n}+\frac{z^2}{4n^2}}}{1+\frac{z^2}{n}}, \tag{3}$$

where CI_l stands for the left extreme of the interval, CI_r for the right one, z is the value of a standard normal distribution at the significant level α (which in this paper was chosen to be $\alpha = 95$ percent), and p is the percentage of success defined by

$$p = \frac{m}{n} = \frac{a_{w,c}}{a_{w,c}+b_{w,c}}. \tag{4}$$

Then, IL is defined by the degree of overlapping between the rule confidence interval and the correspondent *canonical rule* confidence interval:

$$IL(w \to c) = \frac{CI_r(\to c) - CI_l(w \to c)}{CI_r(w \to c) - CI_l(w \to c)}. \quad (5)$$

On the other hand, for selecting a term w, it can be helpful to measure the number of documents of a category c in which w does not appear, that is, to consider the value $c_{w,c}$ (for instance, in [9], it is shown that this helps in classifying the Reuters collection, although is has less impact on Ohsumed). Hence, we obtain L_{ir} (respectively, D_{ir}) from adding the parameter c in (1) (respectively, (2)), with the subindex ir standing for Information Retrieval. The new measures are the following:

$$\begin{aligned} L_{ir}(w \to c) &= \frac{a_{w,c} + 1}{a_{w,c} + b_{w,c} + c_{w,c} + 2}, \\ D_{ir}(w \to c) &= a_{w,c} - b_{w,c} - c_{w,c}. \end{aligned} \quad (6)$$

Finally, considering as p

$$p_{ir}(w \to c) = \frac{a_{w,c}}{a_{w,c} + b_{w,c} + c_{w,c}} \quad (7)$$

in (3), the measure IL_{ir} is obtained.

We could also add to D, L, and IL the parameter $d_{w,c}$. However, the resulting measures behave as D in the sense that the ordering of the words they produce is exactly the same.

3 Linear Filtering Measures

If a category is fixed and the notation given in Section 2.3 is adopted, many of the measures defined in Section 2 accept simple expressions. For instance, if we denote by N the number of documents in the category c and by M the number of documents not in c, we can write

$$\begin{aligned} L(w) &= \frac{a_w + 1}{a_w + b_w + 2}, \\ D(w) &= a_w - b_w, \\ df(w) &= a_w + b_w, \\ CET(w) &= \frac{a_w}{N + M} \log\left(\frac{a_w(N + M)}{N(a_w + b_w)}\right), \end{aligned}$$

where $a_w = a_{w,c}$ and $b_w = b_{w,c}$ for simplicity.

Let us identify each word w with the pair of integer numbers (a_w, b_w). Then, we can study the words that receive identical score under a filtering measure which depends only on (a_w, b_w), by means of the level curves defined by the measure (when it is considered as a bidimensional surface).

In two of the cases above (D and df), the calculation of the level curves is extremely easy. In fact, we have that

$$\begin{aligned} a_w &= b_w + D(w), \\ a_w &= -b_w + df(w), \end{aligned}$$

and, consequently, the level curves of D are straight lines with slope equal to 1 and those of df are also straight lines but with slope -1.

We are interested only in points of the form (a_w, b_w) with a_w and b_w integer (since they represent occurrences of words), and every such point can be conveniently considered isolated from the rest. Thus, this situation can be made general, as the following theorem states (the proofs of the theorems can be found in Appendix A).

Theorem 1. *If m is a filtering measure and N and M are natural numbers, then the level curves passing through the words with $a_w \leq N$ and $b_w \leq M$ can be considered as straight lines.*

However, it is not the case that every measure whose level curves are straight lines has exactly one for each value that it attains. For instance, the measure

$$m(w) = |a_w - b_w|$$

has for each integer value $k \neq 0$ the two level curves

$$\begin{aligned} a_w &= b_w - k, \\ a_w &= b_w + k. \end{aligned}$$

Hence, in general, for each value of a measure can exist several nonintersecting level curves.

Then, an interesting special case is that of measures m which have just one level curve for each value. That is, the measures m that satisfy

$$a_w = f(m(w))b_w + g(m(w))$$

for some functions f and g. For instance, D (with $f(D) = 1$ and $g(D) = D$), L (with $f(L) = \frac{L}{1-L}$ and $g(L) = \frac{2L-1}{1-L}$) and df (with $f(df) = -1$ and $g(df) = df$) are measures with this property.

For these kind of functions, something more can be said, as Theorem 2 shows (see Appendix A for the proof).

Theorem 2. *Suppose N and M are two natural numbers and m is a filtering measure which has exactly one straight line as level curve for each value that m attains over the words with $a_w \leq N$ and $b_w \leq M$. Then, there exist p and q, two polynomials such that*

$$a_w = p(m(w))b_w + q(m(w))$$

for any word w such that $a_w \leq N$ and $b_w \leq M$.

In the light of these two theorems, it is interesting to study which are the filtering measures (and which is their behavior in FS) satisfying

$$a_w = p(m(w))b_w + q(m(w))$$

at least when the degree of p and q is low. In the most simple case, the family of measures obtained is described below (notice that it is not possible that the degrees of p and q are both zero).

When $degree(p) = 0$ and $degree(q) = 1$, we have

$$a_w = c_1 b_w + c_2 m(w) + c_3$$

for some constants c_1, c_2, and c_3 with $c_2 \neq 0$ and, thus,

$$m(w) = \frac{a_w - c_1 b_w - c_3}{c_2}.$$

The value of c_2 can be set to 1, since the ordering produced among the words will always be the same whatever the value of this constant is. Then, we have

$$m(w) = a_w - c_1 b_w - c_3,$$

but now we can make $c_3 = 0$ since, again, the ordering will be unaffected by the value of this constant. Then, we have the parametric family of measures given by

$$m(w) = a_w - c_1 b_w,$$

with c_1 any real number. Some members of this family are D (when $c_1 = 1$), D_{ir} (when $c_1 = \frac{1}{2}$), *df* (when $c_1 = -1$), and the *df* of w in the category c (when $c_1 = 0$).

As we have seen, some of these measures have already been studied in the literature and some have gotten good results in FS for TC. Then, it would be interesting to study the general behavior of this family in the filtering task.

For simplicity, we will work with an equivalent, but slightly modified, expression of the family. We will call *linear filtering measure* to any measure of the form

$$LM_k(w) = ka_w - b_w.$$

These measures have a nice geometrical interpretation which is stated in the following theorem:

Theorem 3. *If w_1 and w_2 are two different words of the same category, then*

$$LM_k(w_1) < LM_k(w_2)$$

if and only if

$$k < \tan\alpha,$$

where α is the angle between the points (b_{w_1}, a_{w_1}) and (b_{w_2}, a_{w_2}).

The family of linear filtering measures depends on the parameter k which can take any real value. However fixed a category, only a finite number of these measures are nonequivalent, in the sense that they produce different orderings of the words (there exist only finitely many such orderings). In fact, the following theorem shows the expression of the number of measures that should be taken into account.

Theorem 4. *The number of nonequivalent filtering measures of the form $ka_w - b_w$ with $k \geq 0$ over $[0, M] \times [0, N]$ is given by*

$$2k_e(N, M) - 1,$$

where k_e is defined recursively by

$$k_e(1, M) = M + 2,$$
$$k_e(N, M) = k_e(N - 1, M) + \varphi_M(N) \text{ if } N > 1$$

with $\varphi_M(N)$ being the quantity of natural numbers between 1 and M (both included) which are relative prime to N.

Obviously, this number of measures is the maximum possible, attained when, for every (a, b) in $[0, M] \times [0, N]$, there exists at least one word in the category with $a_w = a$ and $b_w = b$. If it is not the case, then the number of nonequivalent measures will be lower.

4 The System

In this section, we describe the details of the system that we will apply for TC. This includes the document preprocessing and the way the classification task is tackled.

4.1 Document Preprocessing

As we have already mentioned in Section 1, among the different views of a document, the *bag of words* [2] model is the most widely adopted in TC. It consists in regarding a document as a sequence of terms assuming their independence and ignoring ordering and text structure.

This representation can use Boolean features indicating whether a specific word occurs in a document or not. In addition, it can use the absolute frequency of a word (*tf*) in order to weight its importance in the document. Another measure for this purpose is $tfidf$, which takes into account the distribution of the words in the documents, or its variant tfc (see [11]), which also considers the different lengths of the documents. In this paper, *tf* is chosen because it is one of the most used [1], [6].

In the document representation, different sets of words can be used. One consists of words that belong to each category isolated from the rest, which is known as local lexicum. On the other hand, the global lexicum considers the words from all categories. In this work, the local lexica are considered, since they offer better results [1] (for instance, with the measure D, which is also a linear measure, the best F_1 obtained with global lexica is 82.16 percent in Reuters and 55.11 percent in Ohsumed, much worse than the values obtained with local lexica, as shown in Sections 6.1 and 6.2).

FS will be applied to these lexica (notice that there exists a different one in each category) by means of the measures introduced in Sections 2 and 3.

Additionally, two kinds of feature reduction are performed. The first one consists of removing the *stop words* because they are useless for the classification. The second one involves mapping words with the same meaning to one morphological, which is known as *stemming* [2]. The Porter algorithm [12] is used in this paper for this purpose. This algorithm strips common terminating strings (suffixes) from words in order to reduce them to their roots or stems. A list of suffixes to be removed is specified together with some conditions (for instance, the minimum length of the remaining stem). When the conditions are met, the suffix is stripped or replaced by another suffix.

4.2 The Classification

The aim of TC consists in finding out whether a document is relevant to a category or not. Typically, the task of assigning a category to a document from a finite set of m categories is commonly converted into m binary problems, each one consisting of determining whether a document belongs to a fixed category or not (*one-against-the-rest*) [13]. This transformation makes possible the use of binary classifiers for the multicategory classification problem [11].

In this paper, the classification is performed using Support Vector Machines (SVM) [14], since they have been shown to perform fast [15] and well [16] in TC. The key of this good performance is that SVM are able to handle many features and to deal well with sparse examples. SVM are universal binary classifiers able to find out linear or nonlinear threshold functions to separate the examples of a certain category from the rest. They are based on the *Structural Minimization Risk* principle from computational learning theory [17]. The idea of structural risk minimization is to find a hypothesis h for which it is guaranteed the lowest true error. The true error of h is the probability that h will make an error on an unseen

and randomly selected text example. In this case, the hypothesis that is sought for is a hyperplane which separates the positive training examples from the negative ones leaving the maximum possible margin.

We adopt this approach with a linear threshold function since most TC problems are linearly separable [11].

5 The Experiments

In this section, we describe the settings of the experiments carried out with the linear filtering measures, including the corpora chosen, the way of selecting the parameter k and the measures applied to evaluate the performance.

5.1 The Corpora

In this section, the corpora used in the experiments are described and analyzed. They are the Reuters-21578 collection and the Ohsumed collection.

5.1.1 Reuters-21578 Collection

The Reuters-21578[1] corpus is a set of economic news published by Reuters in 1987.

We chose the Apté split [13] in test and train. After eliminating some documents which contain no body or no topics, we obtained 9,805 documents of which 7,063 are training documents and 2,742 are test ones.

The distribution of documents into the categories is quite unbalanced. In fact, the relative dispersion of the number of documents of the categories is 336.31 percent in the interval [1, 2709] for training documents and 339.86 percent in [1, 1044] for test documents. In addition, 76.40 percent (in train) and 78.65 percent (in test) of the categories have less than 1 percent of the documents.

The words in the corpus are little scattered, since almost half (49.91 percent) of the words appear in only one category and 16.25 percent in only two categories.

5.1.2 Ohsumed Collection

Ohsumed[2] is a clinically oriented MEDLINE subset formed by 348,566 references of 270 medical journals published between 1987 and 1991.

The stories have been manually classified according to the MeSH structure tree (Medical Subjects Headings) [18].

In this work, we considered the first $20,000$ documents of Ohsumed from 1991 with abstract. The first $10,000$ have been labeled as training documents and the rest as test documents. They have been split into the 23 subcategories of diseases (category C) of MeSH (C1, C2,...,C23) [11].

Again, a statistical analysis of this collection was performed. It was found that the distribution of documents over the categories is much more balanced than in Reuters. In fact, the relative dispersion of the number of documents of the categories is 86.91 percent in the interval [100, 2476] for train and 88.51 percent in the interval [82, 2424] for test. Furthermore, only 4.35 percent in train and 8.70 percent in test of the categories have less than 1 percent of the documents, against about 77 percent in Reuters.

The words in this collection are quite more scattered than in Reuters, since 19.55 percent of the words (in average) appear just in one category (against 49.91 percent in Reuters).

1. It is publicly available at http://www.research.attp.com/lewis/reuters21578.html.

2. It is publicly available at http://trec.nist.gov/data/t9-filtering.

5.2 Selection of k

To apply the filtering linear measures, we have to select the value of the parameter k. Since we are interested in testing the performance of the family, we will use several values of k and also adopt several ways of selecting them.

The first one will consist in using the same value of k for filtering in all the categories. For instance, when we apply D or D_{ir}, we are using this approach since the value of k (1 and 2, respectively) is the same for all the categories. We will call these kind of values *absolute values* and we will study their behavior when k ranges in a certain interval. From the results of the measures with k in that interval, we will decide which new values of k are promising and set a new interval of values for study. We will repeat this process until an optimum (possibly local) is found.

However, from the proof of Theorem 4 (see Appendix A), we know that values of k of the form $\frac{b_w}{a_w}$ with w a word of the collection are important. Thus, the second way of selecting the values of k will take into account these values: For each category c, we will compute $k_w = \frac{b_w}{a_w}$ for every w in the lexicum of c and we will order the values k_w. Then, we will select the deciles of that distribution and study the corresponding linear measures. From the results obtained, we will refine the search (taking into account the centiles) as explained for the absolute values. Notice that, in this latter case, the value of k will be different in each category, so we will call these values the *relative values*.

We will be interested in determining whether the use of absolute or of relative values offers better results.

5.3 Evaluation of the Performance

Two measures of effectiveness commonly used in IR [1] are adopted in this paper. Those are *precision* (P) and *recall* (R). The former quantifies the percentage of documents that are correctly classified as positives (they belong to the category) and the latter quantifies the percentage of positive documents that are correctly classified. Due to the trade-off existing between *precision* and *recall* [1], it makes no sense to consider these parameters on their own. Among the several ways of combining them, F_1 is one of the most popular. It gives equal relevance to both *precision* and *recall* and it is defined by

$$F_1 = \frac{1}{0.5\frac{1}{P} + 0.5\frac{1}{R}}.$$

To compute the global performance over all the categories in this work, it is used *microaveraging* which is based on averaging the F_1 of each category in proportion to its number of documents [1].

6 The Results

In this section, we present the results of the experiments described in Section 5 for both corpora under study.

6.1 Reuters-21578

First, we present the results obtained with absolute values of the parameter k (see Section 5.2) and then those obtained with relative values. In each case, we show the results when different numbers of words are filtered out (from 0 percent to 98 percent).

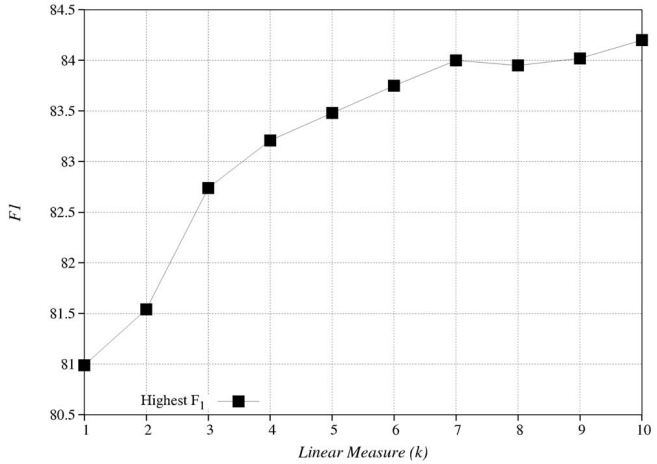

Fig. 1. Highest F_1 on Reuters-21578 of linear measures LM_1 to LM_{10}.

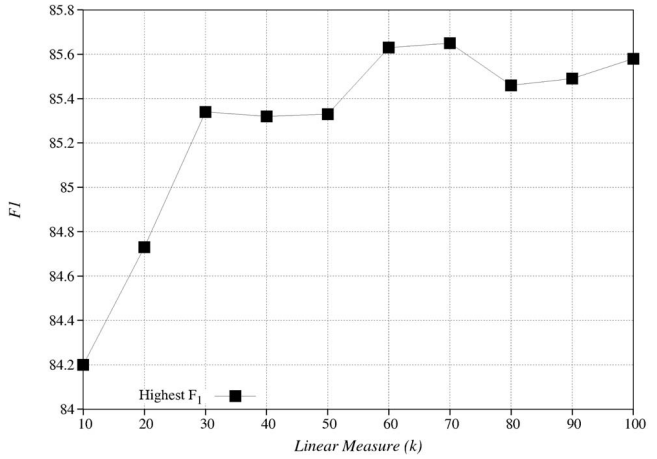

Fig. 2. Highest F_1 on Reuters-21578 of linear measures LM_{10} to LM_{100}.

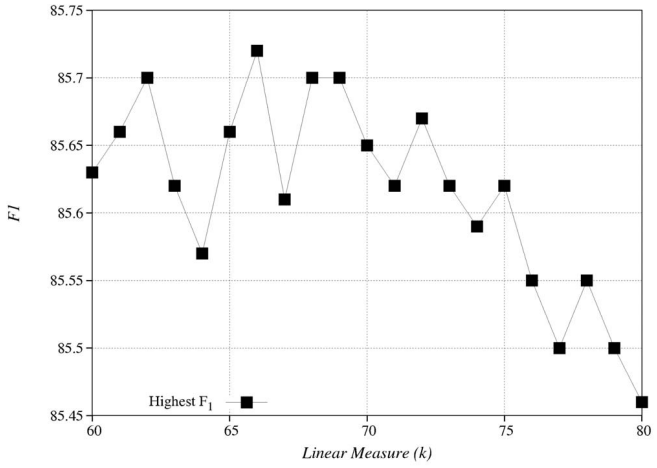

Fig. 3. Highest F_1 on Reuters-21578 of linear measures LM_{60} to LM_{80}.

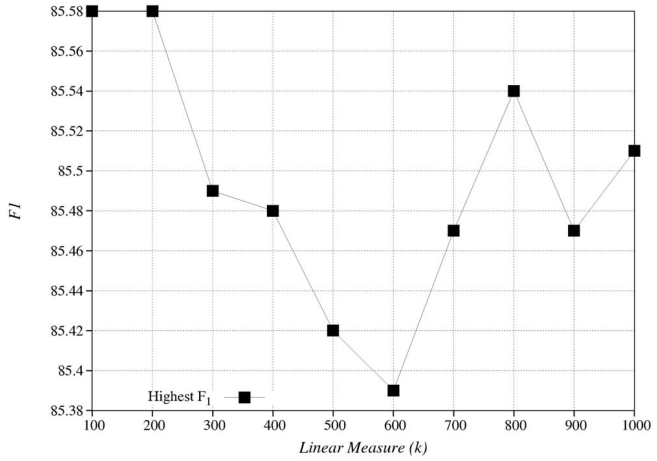

Fig. 4. Highest F_1 on Reuters-21578 of linear measures LM_{100} to LM_{1000}.

6.1.1 Absolute Values

Since $D = LM_1$ and $D_{ir} = LM_2$ (see Section 3), we begin studying the absolute values $k = 1, 2, \ldots, 10$. In Table 1 (all the tables are located in Appendix B which can be found on the Computer Society Digital Library at http://computer.org/tkde/archives.htm), we show the microaveraged F_1 of the classification when the measures $LM_1, \ldots, LM_{10}$ are used for selecting the features (the best value is boldfaced for each measure). The highest values (excepting those obtained when no filtering is applied) attained by each measure are also presented graphically in Fig. 1.

The results obtained with these measures are poor, since no improvement is achieved with regard to no filtering. However, the higher the value of k is, the better the results are (in fact the highest F_1 of this series of experiments is obtained with LM_{10}). Hence, we have performed a new set of experiments, this time with $k = 10, 20, \ldots, 100$. The results are presented in Table 2 and in Fig. 2.

These results are much better, and from LM_{30} on, we obtain measures which can be used for filtering with an improvement of the overall performance for several filtering levels. The best F_1 is achieved by LM_{70}, so we have explored the measures with k between 60 and 70 (see Tables 3 and 4 and Fig. 3).

The highest F_1 is obtained with LM_{66} which attains a value of 85.72 when the 85 percent of the words are filtered.

On the other hand, as Table 2 shows, the measure LM_{100} offers promising results, so it may be interesting to study the behavior of the measures with $k = 100, \ldots, 1,000$. We have conducted experiments similar to the previous ones with these measures and the results are summarized in Table 5 and in Fig. 4. However, none of these measures obtain better results than LM_{66}.

6.1.2 Relative Values

In this section, we present the results of the experiments performed with values of k relative to each category (see Section 5.2).

The first experiments are performed with the values of k corresponding to the deciles of distribution of $\frac{b_w}{a_w}$ in each category. The results are shown in Table 6 and in Fig. 5.

Again, the bigger the value of k, the better results obtained. The highest values of microaveraged F_1 are achieved with the values of k corresponding to the 9th and 10th deciles, so we have performed experiments with the centiles between the 90th and the 100th. The results are shown in Table 7 and in Fig. 6.

The best results are obtained when the value of k is set to the 99th centile.

6.1.3 Comparison with Other Measures

In this section, we compare the best linear measures using absolute and relative values of k (see the two previous

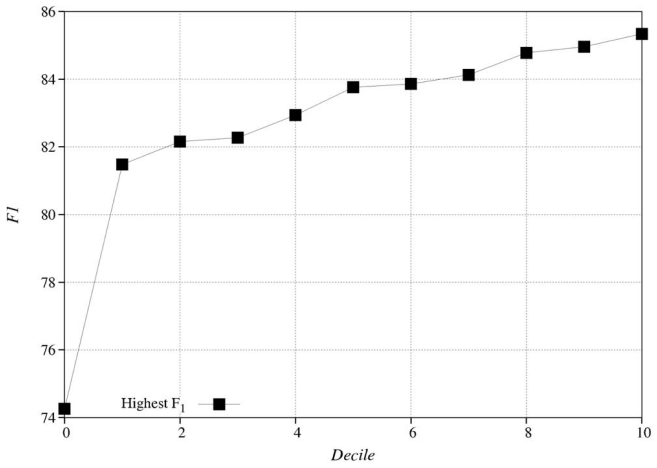

Fig. 5. Highest F_1 on Reuters-21578 of linear measures corresponding to the deciles.

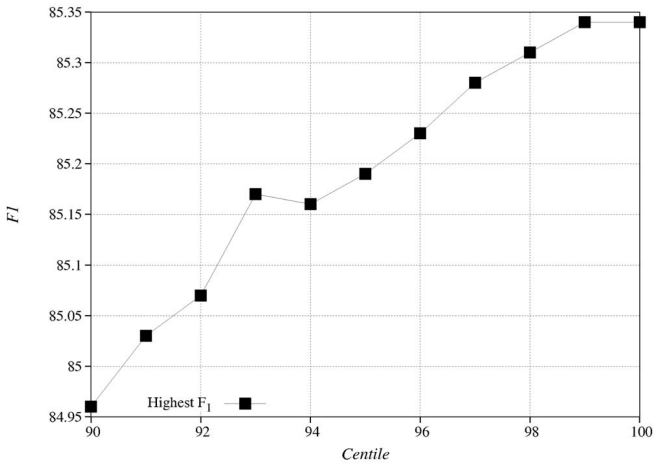

Fig. 6. Highest F_1 on Reuters-21578 of linear measures corresponding to the selected centiles.

sections) with those measures which had obtained the best results in previous studies on FS for TC (see [9]). The results are summarized in Table 8 and graphically presented in Fig. 7. We denote by LM_{c99} the linear measure obtained with values of k corresponding to the 99th centile of the relative values.

The only previously existing measure which outstands the comparison is IG. It is better than LM_{c99}, but LM_{66} obtains the highest results of them all (although IG is more stable when aggressive filtering levels are applied).

6.2 Ohsumed

As with Reuters-21578 (see Section 6.1), we present in first place the results obtained with absolute values of the parameter k and then those obtained with relative ones. Finally, we compare the performance of linear measures with that of the best filtering measures so far.

6.2.1 Absolute Values

Again, we begin studying the absolute values $k = 1, 2, \ldots, 10$, whose microaveraged results of F_1 are shown in Table 9. The highest F_1 achieved by each measure is represented in Fig. 8.

With all these measures, we obtain results that improve the overall performance of the classification in relation to

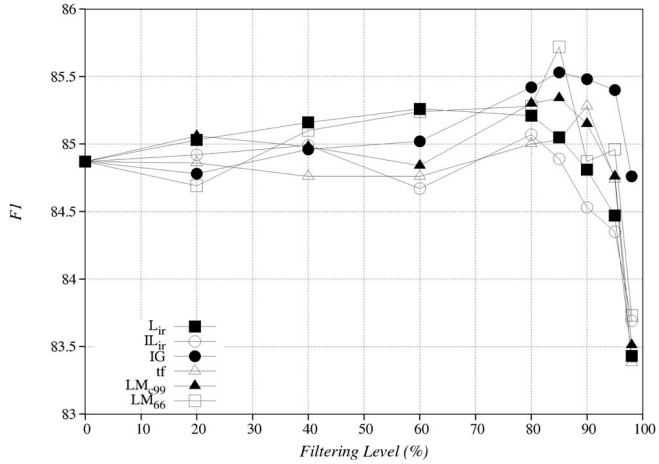

Fig. 7. Comparison of the best measures on Reuters-21578.

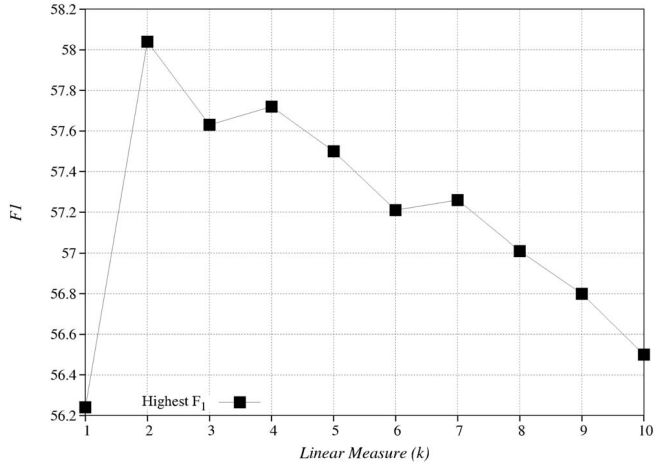

Fig. 8. Highest F_1 on Ohsumed of linear measures LM_1 to LM_{10}.

the case when no filtering is applied. The highest F_1 is achieved by LM_2 (which is D_{ir}), so we explore other measures with values of k near 2.0. The results are shown in Tables 10 and 11 and in Fig. 9.

6.2.2 Relative Values

Now, we present the results of the experiments performed with linear measures whose values of k are relative.

As we did with Reuters-21578, we first experiment with the values of k corresponding to the deciles of $\frac{b_w}{a_w}$ in each category. We present the results in Table 12 and in Fig. 10.

The best values are obtained when k takes the values of the 1st and 2nd decile, so we have performed new experiments with measures whose values of k correspond to the centiles from 10th to 20th. The results are shown in Table 13 and in Fig. 11.

Clearly, the best choice is the measure which corresponds to the 18th centile.

6.2.3 Comparison with Other Measures

As we did in the case of Reuters-21578, this section is devoted to comparing the best linear measures (both with absolute and relative values of the parameter k) to those ones which had offered the best performance in previous studies on FS for TC (see [9]). Table 14 presents the results, where LM_{c18} is the measure corresponding to the

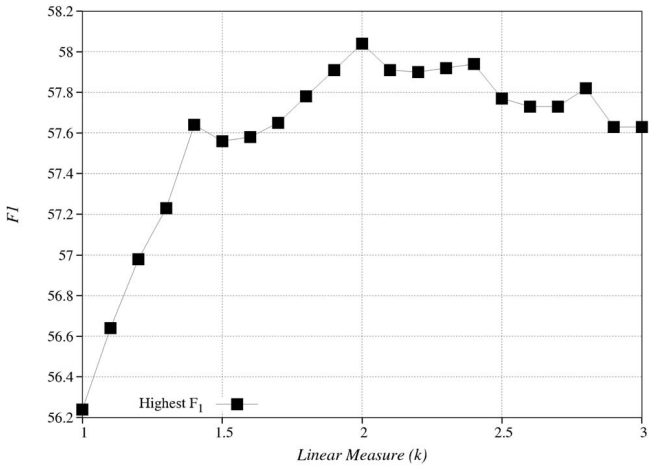

Fig. 9. Highest F_1 on Ohsumed of linear measures $LM_{1.0}$ to $LM_{3.0}$.

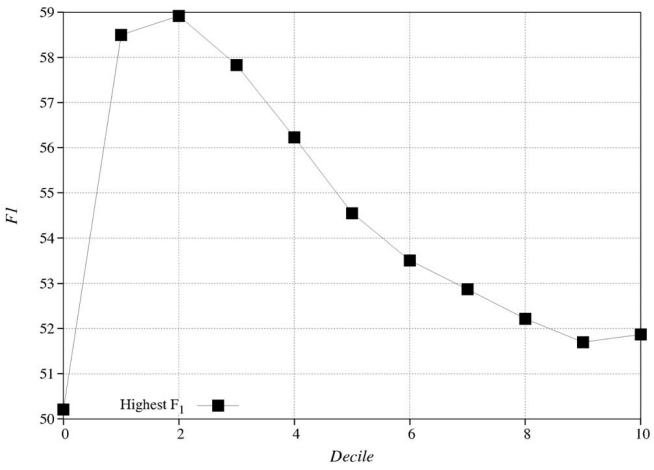

Fig. 10. Highest F_1 on Ohsumed linear measures corresponding to the deciles.

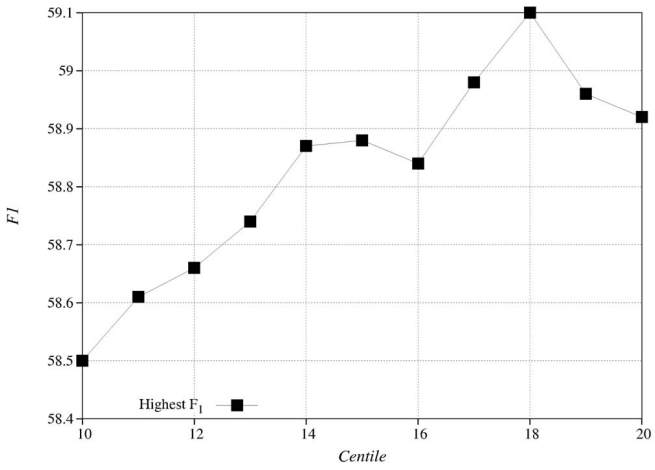

Fig. 11. Highest F_1 on Ohsumed of linear measures corresponding to the selected centiles.

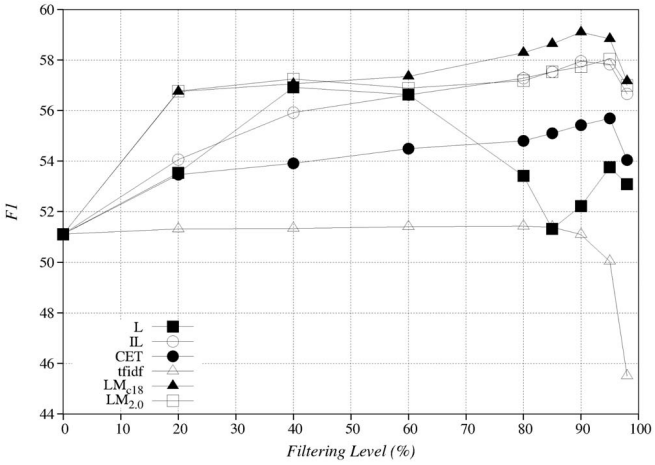

Fig. 12. Comparison of the best measures on Ohsumed.

18th centile. The performance of the measures is graphically represented in Fig. 12.

It is clear from Table 14 and Fig. 12 that both linear measures perform better than any other measure and that LM_{c18} is the best of both. In fact, many of the measures studied in the previous sections also outperform the existing ones on Ohsumed (see Tables 11 and 13). Thus, the family of linear measures seems to be a very appealing choice for performing FS on this collection.

6.3 On the Values of the Parameter k

We have noticed in the previous sections that the linear measures can perform better than previously existing measures both on the Reuters-21578 collection and on the Ohsumed collection. However, the parameters that work well in them are different from one corpus to the other.

While in Reuters-21578, the best choice is a measure with a high absolute value of k (LM_{66}), in Ohsumed, lower values of k chosen relatively in each category seem to be preferable (the best performing measure is LM_{c18}).

These differences may be caused by the different properties of the corpora (see Section 5.1). In Reuters-21578, there exist many words which are specific to a category and hence have low b_w. Thus, the parameter a_w is more discriminative and should be given more relevance (value of k high). On the other hand, in Ohsumed, the words are much more scattered and the importance of a_w and b_w is much more balanced. Consequently, the value of k should not be very high.

7 Conclusions and Further Research

We have introduced a new family of filtering measures for FS in TC which have a simple expression in terms of the appearance of the word in the different documents. We have also shown that these measures have interesting properties (for instance, an appealing geometrical interpretation).

The experiments conducted show that these measures perform at least as well as previously existing measures and, even better, on corpora with different properties (we have used Reuters-21578 and Ohsumed). In fact, in the Ohsumed collection, many linear measures clearly outperform the results obtained with previously existing measures and, in Reuters-21578, the highest F_1 is also obtained by one of the linear measures.

We have also proposed two different methods (the use of absolute and relative values) for the selection of the parameter k on which the family depends. The former

(relative values) performs better on Ohsumed, while the latter (absolute values) is better for Reuters-21578.

This selection of the parameter k is a critical point in the practical application of these measures. The experiments show that the characteristics of the corpora greatly influence this selection. If the words are little scattered (as in the case of Reuters-21578), the importance of the value a_w is bigger and the best overall performance is obtained with high values of k. On the other hand, if the words are more scattered (for instance in Ohsumed), the values of k should be chosen to be lower. We are interested in exploring further this relationship between the optimum value of k and the characteristics of the corpus. Additionally, we plan to develop methods for the automatical determination of the optimal value of k.

We also want to perform a deeper study of the properties of the linear measures as well as to explore further the performance of this family of measures, carrying out experiments on other corpora of similar characteristics to check whether the trends remain the same.

Finally, we are interested in studying other families of measures which can be expressed in a way similar to that of the linear measures.

APPENDIX A
PROOFS OF THE THEOREMS

Proof of Theorem 1. The proof is straightforward. Just notice that there exists an infinite number of slopes such that for any point $p = (b, a)$ in $[0, M] \times [0, N]$, the line determined by them and the point p does not intersect any other point of the form (b_w, a_w) in $[0, M] \times [0, N]$ (it suffices to consider the slope of the line determined by p and $p' = (b', a + 1)$ with b' big enough). □

Proof of Theorem 2. Suppose that $m_1, \ldots, m_k$ are the values that m achieves over the words under the conditions of the theorem. By hypothesis, to each of these values, it corresponds exactly a straight line. Let $a_w = s_i b_w + c_i$ be the line associated to value m_i. It is a classical result of interpolation theory that there exist polynomials p and q such that $p(m_i) = s_i$ and $q(m_i) = c_i$ and, hence, the result. □

Proof of Theorem 3. Obviously,

$$\begin{aligned} LM_k(w_1) < LM_k(w_2) &\iff ka_{w_1} - b_{w_1} < ka_{w_2} - b_{w_2} \\ &\iff k(a_{w_1} - a_{w_2}) < b_{w_1} - b_{w_2} \\ &\iff k < \frac{b_{w_1} - b_{w_2}}{a_{w_1} - a_{w_2}} \end{aligned}$$

and the result follows. □

Proof of Theorem 4. The main point in the proof consists in showing that $k_e(N, M)$ is the number of measures which assign equal value to at least two points in $[0, M] \times [0, N]$ and that, in fact, this coincides with the number of measures that assign value 0 to the point $(0, 0)$ and, at least, to another different point in $[0, M] \times [0, N]$.

If this is so, then it is clear that the total number of nonequivalent measures will be $2k_e(N, M) - 1$. In fact, if we take two measures which assign the value 0 to at least one point different from $(0, 0)$ and their k's (namely, $k_1 < k_2$) are the closest possible, then all the measures with k between k_1 and k_2 form an equivalence class.

Then, we have to show:

- That any measure which assigns the same value to two different points in $[0, M] \times [0, N]$ in fact assigns the value 0 to a point different in $[0, M] \times [0, N]$ different from $(0, 0)$.
- That the number of such measures is given by $k_e(N, M)$.

To prove the first point, consider a measure $m(w) = ka_w - b_w$ which assigns the same value to the points (b_1, a_1) and (b_2, a_2). We have several cases:

1. If $k = 0$ or $k = \infty$, the result follows immediately.
2. If $k > 0$, then we can consider without loss of generality that $a_1 > a_2$ and then
$$ka_1 - b_1 = ka_2 - b_2$$
and, hence,
$$k(a_1 - a_2) = b_1 - b_2.$$
But, $a_1 - a_2 > 0$ and $k > 0$, so $b_1 - b_2 > 0$ and the point $(a_1 - a_2, b_1 - b_2)$ belongs to $[0, M] \times [0, N]$ is different from $(0, 0)$ and, obviously, attains the value 0.

Now, we have to prove that the number of measures which assign 0 to a point different from $(0, 0)$ is given by

$$\begin{aligned} &k_e(1, M) = M + 2, \\ &k_e(N, M) = k_e(N - 1, M) + \varphi_M(N) \text{ if } N > 1. \end{aligned}$$

We proceed by induction on N. If $N = 1$, then there exists exactly $M + 2$ two different measures of that kind, $M + 1$ corresponding to each straight line from $(0, 0)$ to the points $(0, 1), (1, 1), (2, 1), \ldots, (M, 1)$ and the other one corresponding to the straight line from $(0, 0)$ to $(M, 0)$.

If $N > 1$, then we have to sum up the lines assigning 0 to points of the form (i, j) with $0 \le i \le M$ and $0 \le j < N$, which is given by $k_e(N - 1, M)$ (induction hypothesis) and the lines assigning 0 to points of the form (i, N) with $1 \le i \le M$ which are not included in the lines counted by $k_e(N - 1, M)$ (notice that the line joining $(0, 0)$ and $(0, N)$ is already included).

But, a line $ka - b$ which assigns 0 to a point (i, N) also assigns 0 to a point $(b, a) \neq (0, 0)$ with $a < N$ if and only if

$$ka - b = 0 = kN - i.$$

Then, we can deduce that

$$k = \frac{i}{N} = \frac{b}{a}$$

and, thus,

$$ai = bN.$$

Then, ai is divided by i and N and, hence, d, the least common multiple of i and N is less than or equal to ai. But, since $a < N$, it follows that $d \le ai < iN$ and, thus, i and N are not relative primes.

Similarly, it is easy to prove that if i and N are not relative primes, then there exists $(a, b) \neq (0, 0)$ with $a < N$ and $b \leq M$ such that $ai = bN$. Then, the line joining $(0, 0)$ and (i, N) is already counted in $k_e(N - 1, M)$.

Consequently, a new line should be counted if and only if i and N are relative primes and the result follows. □

ACKNOWLEDGMENTS

This research been supported in part under MCyT and Feder grant TIN2004-05920.

REFERENCES

[1] F. Sebastiani, "Machine Learning in Automated Text Categorisation," *ACM Computing Survey*, vol. 34, no. 1, 2002.
[2] G. Salton and M.J. McGill, *An Introduction to Modern Information Retrieval*. McGraw-Hill, 1983.
[3] G.H. John, R. Kohavi, and K. Pfleger, "Irrelevant Features and the Subset Selection Problem," *Proc. 11th Int'l Conf. Machine Learning*, pp. 121-129, 1994.
[4] Y. Yang and J.O. Pedersen, "A Comparative Study on Feature Selection in Text Categorisation," *Proc. 14th Int'l Conf. Machine Learning*, pp. 412-420, 1997.
[5] I. Díaz, J. Ranilla, E. Montañés, J. Fernández, and E.F. Combarro, "Improving Performance of Text Categorisation by Combining Filtering and Support Vector," *J. Am. Soc. Information Science and Technology (JASIST)*, vol. 55, no. 7, pp. 579-592, 2004.
[6] D. Mladenic and M. Grobelnik, "Feature Selection for Unbalanced Class Distribution and Naive Bayes," *Proc. 16th Int'l Conf. Machine Learning*, pp. 258-267, 1999.
[7] L. Galavotti, F. Sebastiani, and M. Simi, "Experiments on the Use of Feature Selection and Negative Evidence in Automated Text Categorization," *Proc. Fourth European Conf. Research and Advanced Technology for Digital Libraries*, pp. 59-68, 2000.
[8] E. Montañés, J. Fernández, I. Díaz, E.F. Combarro, and J. Ranilla, "Measures of Rule Quality for Feature Selection in Text Categorization," *Proc. Fifth Int'l Symp. Intelligent Data Analysis Berlin*, vol. 2810, pp. 589-598, 2003.
[9] E. Montañés, I. Díaz, J. Ranilla, E. Combarro, and J. Fernández, "Scoring and Selecting Terms for Text Categorization," *IEEE Intelligent Systems*, to appear.
[10] J. Fürnkranz and G. Widmer, "Incremental Reduced Error Pruning," *Proc. Int'l Conf. Machine Learning*, pp. 70-77, citeseer.ist.psu.edu/furnkranz94incremental.html, 1994.
[11] T. Joachims, "Text Categorization with Support Vector Machines: Learning with Many Relevant Features," *Proc. 10th European Conf. Machine Learning*, no. 1398, pp. 137-142, 1998.
[12] M.F. Porter, "An Algorithm for Suffix Stripping," *Program (Automated Library and Information Systems)*, vol. 14, no. 3, pp. 130-137, 1980.
[13] C. Apte, F. Damerau, and S. Weiss, "Automated Learning of Decision Rules for Text Categorization," *Information Systems*, vol. 12, no. 3, pp. 233-251, 1994.
[14] C. Cortes and V. Vapnik, "Support-Vector Networks," *Machine Learning*, vol. 20, no. 3, pp. 273-297, 1995.
[15] S. Dumais, J. Platt, D. Heckerman, and M. Sahami, "Inductive Learning Algorithms and Representations for Text Categorization," *Proc. Int'l Conf. Information and Knowledge Management*, pp. 148-155, 1998.
[16] Y. Yang and X. Liu, "A Re-Examination of Text Categorization Methods," *Proc. 22nd ACM Int'l Conf. Research and Development in Information Retrieval*, pp. 42-49, citeseer.nj.nec.com/yang99reexamination.html, 1999.
[17] V. Vapnik, *The Nature of Statistical Learning Theory*. Springer-Verlag, 1995.
[18] National Library of Medicine, "Medical Subject Headings (Mesh)," http://www.nlm.nih.gov/mesh/2002/index.html, 1993.

Elías F. Combarro received the PhD degree in mathematics from the University of Oviedo. He is an assistant professor of computer science at the University of Oviedo, Spain. Formerly, he conducted research in computability theory, studying the symmetry of some computable objects, like Kleene's universal function. Nowadays, his research is focused in artificial intelligence, machine learning, information retrieval, and text categorization.

Elena Montañés received the PhD degree in computer science and artificial intelligence from the University of Oviedo. She is an assistant professor of computer science and artificial intelligence at Oviedo University in Spain. Her research interests focus on text categorization, machine learning, artificial intelligence, feature selection, and time series forecasting.

Irene Díaz received a degree in mathematics from the University of Oviedo and the PhD degree in computer science from the University Carlos III of Madrid. She is an assistant professor of computer science at the University of Oviedo in Spain. Her research interests include information retrieval, knowledge management, parallel computing, and artificial intelligence.

José Ranilla received a degree in computer science from the Technical University of Valencia and the PhD degree in computer science from the University of Oviedo. He is an assistant professor of computer science at the University of Oviedo in Spain. His research interests include information retrieval, knowledge management, parallel computing, and machine learning.

Ricardo Mones is working as a researcher in the Information Retrieval Group of the Artificial Intelligence Center of the University of Oviedo. His research interests include the use of self-organizing maps in information retrieval, text categorization, natural language processing, and artificial intelligence.