Supervised Term Weighting for Automated Text Categorization

Franca Debole
Istituto di Scienza e Tecnologie dell'Informazione
Consiglio Nazionale delle Ricerche
Via G. Moruzzi 1 - 56124 Pisa (Italy)
debole@iei.pi.cnr.it

Fabrizio Sebastiani
Istituto di Scienza e Tecnologie dell'Informazione
Consiglio Nazionale delle Ricerche
Via G. Moruzzi 1 - 56124 Pisa (Italy)
fabrizio@iei.pi.cnr.it

ABSTRACT

The construction of a text classifier usually involves (i) a phase of *term selection*, in which the most relevant terms for the classification task are identified, (ii) a phase of *term weighting*, in which document weights for the selected terms are computed, and (iii) a phase of *classifier learning*, in which a classifier is generated from the weighted representations of the training documents. This process involves an activity of *supervised learning*, in which information on the membership of training documents in categories is used. Traditionally, supervised learning enters only phases (i) and (iii). In this paper we propose instead that learning from training data should also affect phase (ii), i.e. that information on the membership of training documents to categories be used to determine term weights. We call this idea *supervised term weighting* (STW). As an example, we propose a number of "supervised variants" of $tfidf$ weighting, obtained by replacing the *idf* function with the function that has been used in phase (i) for term selection. We present experimental results obtained on the standard Reuters-21578 benchmark with one classifier learning method (support vector machines), three term selection functions (information gain, chi-square, and gain ratio), and both local and global term selection and weighting.

Keywords

Machine learning, text categorization, text classification

1. INTRODUCTION

Text categorization (TC) is the activity of automatically building, by means of machine learning (ML) techniques, *automatic text classifiers*, i.e. programs capable of labelling natural language texts from a domain $\mathcal{D}$ with thematic categories from a predefined set $\mathcal{C} = \{c_1, \ldots, c_{|\mathcal{C}|}\}$ [10]. The construction of an automatic text classifier relies on the existence of an *initial corpus* $\Omega = \{d_1, \ldots, d_{|\Omega|}\}$ of documents preclassified under $\mathcal{C}$. A general inductive process (called the *learner*) automatically builds a classifier for $\mathcal{C}$ by learning the characteristics of $\mathcal{C}$ from a *training set* $Tr = \{d_1, \ldots, d_{|Tr|}\}$ of documents. Once a classifier has been built, its effectiveness (i.e. its capability to take the right categorization decisions) may be tested by applying it to the *test set* $Te = \Omega - Tr$ and checking the degree of correspondence between the decisions of the classifier and those encoded in the corpus. This is called a *supervised* learning activity, since learning is "supervised" by the information on the membership of training documents in categories.

The construction of a text classifier may be seen as consisting of essentially two phases:

1. *document indexing*, i.e. the creation of internal representations for documents. This typically consists in
   (a) *term selection*, consisting in the selection, from the set $\mathcal{T}$ (that contains of all the terms that occur in the documents of Tr), of the subset $\mathcal{T}' \subset \mathcal{T}$ of terms that, when used as dimensions for document representation, are expected to yield the best effectiveness; and
   (b) *term weighting*, in which, for every term t_k selected in phase (1a) and for every document d_j, a weight $0 \leq w_{kj} \leq 1$ is computed which represents, loosely speaking, how much term t_k contributes to the discriminative semantics of document d_j;
2. a phase of *classifier learning*, i.e. the creation of a classifier by learning from the internal representations of the training documents.

Traditionally, supervised learning affects only phases (1a) and (2). In this paper we propose instead that supervised learning is used also in phase (1b), so as to make the weight w_{kj} reflect the importance of term t_k in deciding the membership of d_j to the categories of interest. We call this idea *supervised term weighting* (STW).

Concerning the computation of term weights, we propose that phase (1b) capitalizes on the results of phase (1a), since the selection of the best terms is usually accomplished by scoring each term t_k by means of a function $f(t_k, c_i)$ that measures its capability to discriminate category c_i, and then selecting the terms that maximize $f(t_k, c_i)$. In our proposal the $f(t_k, c_i)$ scores are not discarded after term selection, but become an active ingredient of the term weight.

The TC literature discusses two main policies to perform term selection: (a) a *local* policy, where different sets of

SAC 2003, Melbourne, Florida USA

terms $\mathcal{T}_i' \subset \mathcal{T}$ are selected for different categories c_i, and (b) a *global* policy, where a single set of terms $\mathcal{T}' \subset \mathcal{T}$ is selected by extracting a single score $f_{glob}(t_k)$ from the individual scores $f(t_k, c_i)$. In this paper we experiment with both policies, but always using the same policy for both term selection and term weighting. A consequence of adopting the local policy and reusing the scores for term weighting is that weights, traditionally a function of a term t_k and a document d_j, now also depend on a category c_i; this means that, in principle, the representation of a document is no more a vector of $|\mathcal{T}'|$ terms, but a set of vectors of $\mathcal{T}_i'$ terms, with $i = 1, \ldots, |\mathcal{C}|$.

The paper is organized as follows. Section 2 discusses the roles that term selection and term weighting play in current approaches to TC. In Section 3 we describe in detail the idea behind STW, and introduce some example weighting functions based on this idea. In Section 4 we experiment these functions on Reuters-21578, the standard benchmark of TC research. Experiments have been performed with one classifier learning method (support vector machines), three term selection functions (information gain, chi-square, and gain ratio), and both local and global term selection and weighting. Section 5 concludes.

2. DOCUMENT INDEXING IN TC

2.1 Term weighting

In text categorization and other applications at the crossroads of IR and ML, term weighting is usually tackled by means of methods borrowed from text search, i.e. methods that do not involve a learning phase. Many weighting methods have been developed within text search, and their variety is astounding. However, as noted by Zobel and Moffat [13] (from which the passages below are quoted), there are three *monotonicity assumptions* that, in one form or another, appear in practically all weighting methods: (i) "rare terms are no less important than frequent terms" (the *IDF assumption*); (ii) "multiple appearances of a term in a document are no less important than single appearances" (the *TF assumption*); (iii) "for the same quantity of term matching, long documents are no more important than short documents" (the *normalization assumption*). These assumptions are well exemplified by the $tfidf$ function (here presented in its standard "ltc" variant [9]), i.e.

$$tfidf(t_k, d_j) = tf(t_k, d_j) \cdot \log \frac{|Tr|}{\#_{Tr}(t_k)} \quad (1)$$

where $\#_{Tr}(t_k)$ denotes the number of documents in Tr in which t_k occurs at least once and

$$tf(t_k, d_j) = \begin{cases} 1 + \log \#(t_k, d_j) & \text{if } \#(t_k, d_j) > 0 \\ 0 & \text{otherwise} \end{cases}$$

where $\#(t_k, d_j)$ denotes the number of times t_k occurs in d_j. The $tf(t_k, d_j)$ and $\log \frac{|Tr|}{\#_{Tr}(t_k)}$ components of Equation (1) enforce the TF and IDF assumptions, respectively. Weights obtained by Equation (1) are usually normalized by cosine normalization, i.e.

$$w_{kj} = \frac{tfidf(t_k, d_j)}{\sqrt{\sum_{s=1}^{|\mathcal{T}|} tfidf(t_s, d_j)^2}} \quad (2)$$

which enforces the normalization assumption.

2.2 Term selection

Many classifier induction methods are computationally hard, and their computational cost is a function of the length of the vectors that represent the documents. It is thus of key importance to be able to work with vectors shorter than $|\mathcal{T}|$, which is usually a number in the tens of thousands or more. For this, *term selection* techniques are used to select from $\mathcal{T}$ a subset $\mathcal{T}'$ (with $|\mathcal{T}'| \ll |\mathcal{T}|$) of terms that are deemed most useful for compactly representing the meaning of the documents. The value

$$\xi = \frac{|\mathcal{T}| - |\mathcal{T}'|}{|\mathcal{T}|} \quad (3)$$

is called the *reduction factor*. Usually, these techniques consist in scoring each term in $\mathcal{T}$ by means of a category-based *term evaluation function* f (TEF) and then selecting a set $\mathcal{T}'$ of terms that maximize f. Many functions, mostly from the tradition of information theory and statistics, have been used as TEFs in TC; those of interest to the present work are

$$\chi^2(t_k, c_i) = \frac{[P(t_k, c_i)P(\bar{t}_k, \bar{c}_i) - P(t_k, \bar{c}_i)P(\bar{t}_k, c_i)]^2}{P(t_k)P(\bar{t}_k)P(c_i)P(\bar{c}_i)} \quad (4)$$

$$IG(t_k, c_i) = \sum_{c \in \{c_i, \bar{c}_i\}} \sum_{t \in \{t_k, \bar{t}_k\}} P(t, c) \log_2 \frac{P(t, c)}{P(t)P(c)} \quad (5)$$

$$GR(t_k, c_i) = \frac{\sum_{c \in \{c_i, \bar{c}_i\}} \sum_{t \in \{t_k, \bar{t}_k\}} P(t, c) \log_2 \frac{P(t, c)}{P(t)P(c)}}{-\sum_{c \in \{c_i, \bar{c}_i\}} P(c) \log_2 P(c)} \quad (6)$$

which are called *chi-square*, *information gain*, and *gain ratio*, respectively. In these formulae, probabilities are interpreted on an event space of documents (e.g. $P(\bar{t}_k, c_i)$ indicates the probability that, for a random document x, term t_k does not occur in x and x belongs to category c_i), and are estimated by maximum likelihood.

All of these functions try to capture the intuition according to which the most valuable terms for categorization under c_i are those that are distributed most differently in the sets of positive and negative examples of c_i. However, interpretations of this principle may vary subtly across different functions; see Section 4.1 for a discussion.

Equations (4), (5) and (6) refer to a specific category c_i; in order to assess the value of a term t_k in a "global", category-independent sense, a "globalization" technique is applied so as to extract a global score $f_{glob}(t_k)$ from the $f(t_k, c_i)$ scores relative to the individual categories. The most common globalization techniques are the sum $f_{sum}(t_k) = \sum_{i=1}^{|\mathcal{C}|} f(t_k, c_i)$, the weighted sum $f_{wsum}(t_k) = \sum_{i=1}^{|\mathcal{C}|} P(c_i) f(t_k, c_i)$, and the maximum $f_{max}(t_k) = \max_{i=1}^{|\mathcal{C}|} f(t_k, c_i)$ of their category-specific values $f(t_k, c_i)$.

3. SUPERVISED TERM WEIGHTING

While the normalized $tfidf$ function of Equation (2), or other term weighting functions from the IR literature, are routinely used in IR applications involving supervised learning such as TC, we think that their use in these contexts is far from being the optimal choice. In particular, the present paper challenges the IDF assumption. In standard IR contexts this assumption is reasonable, since it encodes the quite plausible intuition that a term t_k that occurs in too

many documents is not sufficiently helpful, when it occurs in a query q, in discriminating the documents relevant to q from the irrelevant. However, if training data for the query were available (i.e. documents whose relevance or irrelevance to q is known), an even stronger intuition should be brought to bear, i.e. the one according to which the best discriminators are the terms that are distributed most differently in the sets of positive and negative training examples.

Training data is not available for queries in standard IR contexts, but is available for categories in TC contexts. In these contexts, *category-based* TEFs (such as (4), (5) and (6)) that score terms according to how differently they are distributed in the sets of positive and negative training examples, are thus better substitutes of *idf*-like functions.

An attractive aspect of using STW in TC is that, when category-based TEFs have been used for term selection, the scores they attribute to terms are already available. Therefore, the approach we propose here puts the scores computed in the phase of term selection to maximum use: instead of discarding them after selecting the terms that will take part in the representations, they are used also in the term weighting phase.

4. EXPERIMENTS

We have conducted a number of experiments to test the validity of the STW idea. The experiments have been run on a standard benchmark using three different TEFs, employed both according to the local and global policies, and always using the same TEF both as the term selection function *and* as a component of the term weighting function. Therefore, when we speak e.g. of using $IG(g)$ as a STW technique, we mean using IG (according to the global policy, denoted by "(g)" – local is denoted by "(l)") both as a term selection function *and* as a substitute of $\log \frac{|Tr|}{\#_{Tr}(t_k)}$ in Equations (1) and (2).

4.1 Term evaluation functions

In our experiments we have used the three TEFs illustrated in Equations (4), (5) and (6). The first two have been chosen since they are the two most frequently used category-based TEFs in the TC literature (document frequency is also often used as a TEF [12], but it is not category-based), while the third has been chosen since, as we discuss below, we consider it a theoretically better motivated variant of the second.

The first TEF we discuss is the chi-square (χ^2) statistics, which is frequently used in the experimental sciences in order to measure how the results of an observation differ (i.e. are independent) from the results expected according to an initial hypothesis (lower values indicate lower dependence)[1]. In term selection we measure how independent t_k and c_i are. The terms t_k with the lowest value for $\chi^2(t_k, c_i)$ are thus the most independent from c_i; since we are interested in the terms which are not, we select the terms t_k for which $\chi^2(t_k, c_i)$ is highest.

The second TEF we employ is *information gain* (IG), an information-theoretic function which measures the amount of information one random variable contains about another (or, in other words, the reduction in the uncertainty of a random variable that knowledge of the other brings about)[2]; it is 0 for two independent variables, and grows monotonically with their dependence [1]. In term selection we measure how much information term t_k contains about category c_i, and we are interested in selecting the terms that are more informative about (i.e. more indicative of the presence or of the absence of) the category, so we select the terms for which $IG(t_k, c_i)$ is highest.

The third TEF we discuss is *gain ratio* (GR), defined as the ratio between the information gain $IG(X, Y)$ of the two variables X and Y and the entropy of one of them ($H(X)$ or $H(Y)$) [8]. Although, to our knowledge, GR has never been used for feature selection purposes, we claim that for term selection it is a better alternative than IG since, as Manning and Schütze [7, p. 67] note, IG grows not only with the degree of dependence of the two variables, but also with their entropy. Dividing $IG(t_k, c_i)$ by $H(c_i) = -\sum_{c \in \{c_i, \bar{c}_i\}} P(c) \log_2 P(c)$ allows us to compare the different values of term t_k for different categories on an equal basis. Note in fact that while $0 \leq IG(t_k, c_i) \leq min\{H(t_k), H(c_i)\}$, we have instead that $0 \leq GR(t_k, c_i) \leq 1$. Comparing the different scores that t_k has obtained on the different categories is especially important when applying the globalization techniques described in Section 2.2. For instance, it is clear that if we choose IG as our TEF and $f_{max}(t_k) = \max_{i=1}^{|\mathcal{C}|} f(t_k, c_i)$ as our globalization function, the score $IG(t_k, c_1)$ for a category c_1 with high entropy has a higher probability of being selected that the score $IG(t_k, c_2)$ for a category c_2 with low entropy. Instead, with GR these categories do not enjoy this "unfair advantage".

4.2 Learning method

Since a document d_j can belong to zero, one or many of the categories in $\mathcal{C}$, we tackle the classification problem as $|\mathcal{C}|$ independent problems of deciding whether d_j belongs or not to c_i, for $i = 1, \ldots, |\mathcal{C}|$.

The learning method used for our experiments is a support vector machine (SVM) learner as implemented in the SVM-LIGHT package (version 3.5) [4]. SVMs attempt to learn a hyperplane in $|\mathcal{T}|$-dimensional space that separates the positive training examples from the negative ones with the maximum possible margin, i.e. such that the minimal distance between the hyperplane and a training example is maximum; results in computational learning theory indicate that this tends to minimize the generalization error, i.e. the error of the resulting classifier on yet unseen examples. We have simply opted for the default parameter setting of SVMLIGHT; in particular, this means that a linear kernel has been used.

In an extended version of this paper [2] we also discuss analogous experiments we have carried out with two other learners (a Rocchio method and k-NN algorithm), and with three different reduction factors (.00, .50, .90).

4.3 Experimental setting

In our experiments we have used the "Reuters-21578, Distribution 1.0" corpus, currently the most widely used benchmark in TC research[3]. Reuters-21578 consists of a set of

[1]Since χ^2 is a statistics, it is usually best viewed in terms of actual counts from a contingency table, and not in terms of probabilities. In (4) we have formulated χ^2 in probabilistic terms for better comparability with the other two TEFs.

[2]Information gain is also known as *mutual information* [7, pp. 66 and 583]. Although many TC researchers have used this function under one name or the other, the fact that the two names refer to the same object seems to have gone undetected.

[3]The Reuters-21578 corpus is freely avail-

12,902 news stories, partitioned (according to the "ModApté" split we have adopted) into a training set of 9,603 documents and a test set of 3,299 documents. The documents are labelled by 118 categories; the average number of categories per document is 1.08, ranging from a minimum of 0 to a maximum of 16. The number of positive examples per category ranges from a minimum of 1 to a maximum of 3964.

All our results are reported (a) for the set of 115 categories with at least one training example (hereafter, Reuters-21578(115)), (b) for the set of 90 categories with at least one training example and one test example (Reuters-21578(90)), and (c) for the set of the 10 categories with the highest number of training examples (Reuters-21578(10)). Sets (a) and (b) are obviously the hardest, since they include categories with very few positive instances for which inducing reliable classifiers is obviously a haphazard task.

In all the experiments discussed in this section, stop words have been removed using the stop list provided in [5, pages 117–118]. Punctuation has been removed, all letters have been converted to lowercase, numbers have been removed, and stemming has been performed by means of Porter's stemmer. We have measured effectiveness in terms of precision wrt c_i (π_i) and recall wrt c_i (ρ_i), defined in the usual way. Values relative to individual categories are averaged to obtain values of precision (π) and recall (ρ) global to the entire category set according to the two alternative methods of microaveraging and macroaveraging. Note that for the computation of macroaveraging, conforming to common practice, we have taken π_i (resp. ρ_i) to be 1 when the denominator $TP_i + FP_i$ (resp. $TP_i + FN_i$) is 0.

As a measure of effectiveness that combines the contributions of π and ρ we have used the well-known F_β function [6], defined as

$$F_\beta = \frac{(\beta^2 + 1)\pi\rho}{\beta^2\pi + \rho}$$

with $0 \leq \beta \leq +\infty$. Similarly to most other researchers we have set $\beta = 1$, which places equal emphasis on π and ρ. The results of our experiments are reported in Figure 1.

Whenever term selection has been performed according to the global policy, the $f_{max}(t_k)$ has been used as the globalization technique, since in preliminary experiments we have run it consistently outperformed the other globalization techniques described in Section 2.2. The reason why $f_{max}(t_k)$ performs well is that it prefers terms that are very good separators even on a single category, rather than terms that are only "fair" separators on many categories. In fact, if t_k is a very good separator for c_i, then $f(t_k, c_i)$ is going to be very high, so that there are good chances that $f_{max}(t_k) = f(t_k, c_i)$, which means that there are good chances that t_k is selected, which means in turn that there is a good separator for c_i in the selected term set.

In all experiments, STW techniques have been compared with a baseline formed by cosine-normalized $tfidf$ weighting (in the "ltc" variant of Equations (1) and (2)). Note that although stronger weighting functions than "ltc" $tfidf$ have been reported in the literature [13], all of them are based on the three monotonicity assumptions mentioned in Section 2.1; this means that our STW techniques could be applied to them too, probably yielding similar performance differentials.

able for experimentation purposes from http://www.daviddlewis.com/resources/testcollections/reuters21578/

4.4 Analysis of the results: STW functions

The thorough experiments we have performed have not shown a uniform superiority of STW with respect to standard term weighting: in some cases $tfidf$ has outperformed all STW techniques, while in other cases some of the STW techniques have improved on $tfidf$. Let us try to analyze the results more in detail; for ease of discussion we will refer to the results obtained on Reuters-21578(90).

Weighting techniques $GR(g)$ and $\chi^2(g)$ are the best performers for SVMs (although on SVMs $tfidf$ is just as good on microaveraging). The fact that both $\chi^2(g)$ and $GR(g)$ have achieved an 11% improvement (.582 vs. .524) on macroaveraged effectiveness over the best $tfidf$ for SVMs, while basically maintaining the same microaveraged effectiveness, is of particular relevance, since SVMs are currently the best performing TC method in the literature. $IG(g)$ is instead a disappointing performer, sometimes disastrously so (namely, with macroaveraging). Among the local policies, $GR(l)$ is again generally the best, with $IG(l)$ usually faring better than $\chi^2(l)$.

We are not surprised by the good performance of $GR(g)$ since, as we have remarked in Section 4.1, we consider $GR(g)$ a theoretically superior alternative to $IG(g)$. The disappointing performance that this latter has produced is a striking contrast with the well-known good performance of IG as a term selection function [12]. Note that $IG(l)$ and $GR(l)$ perform identically. This is due to the fact that the two differ only by the entropy of c_i being used as a normalization factor in $GR(l)$. Therefore, it is quite obvious that, locally to category c_i, $IG(l)$ and $GR(l)$ select the same terms and give them weights that differ only by a constant multiplicative factor.

A surprising result is that global STW techniques are almost everywhere superior to the corresponding local technique. We say this is surprising because the global policy openly contradicts the decision to view the classification problem as $|\mathcal{C}|$ independent binary classification problems. That is, if these $|\mathcal{C}|$ problems are really to be seen as independent, then the problem of building representations for them should also be viewed on a category-by-category basis, which is what the local policy does. We conjecture that this surprising behaviour is due to the fact that the statistics that can be collected from scarcely populated categories are not robust enough for the local policy to be effective, and that for these categories the global policy makes up for their unreliable statistics by providing more robust statistics collected over the entire category set.

5. CONCLUSION

We have proposed *supervised term weighting* (STW), a term weighting methodology specifically designed for IR applications involving supervised learning, such as text categorization and text filtering. Supervised term indexing leverages on the training data by weighting a term according to how different its distribution is in the positive and negative training examples. We have also proposed that this should take the form of replacing idf by the category-based term evaluation function that has previously been used in the term selection phase; as such, STW is also efficient, since it reuses for weighting purposes the scores already computed for term selection purposes.

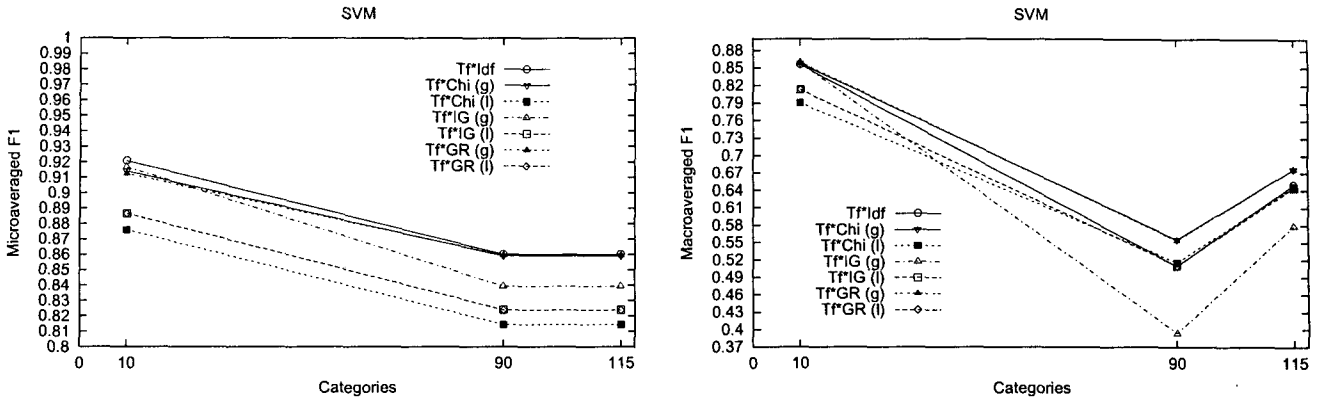

Figure 1: Plots of micro-averaged F_1 (leftmost) and macro-averaged F_1 (rightmost) for SVMs. The X axis indicates the three major subsets of Reuters-21578 described in Section 4.3.

We have tested STW in all the combinations involving one learning methods and three different term weighting functions, each tested in its local and global version. One of these functions (gain ratio) was not known from the TC term selection literature, and was proposed here since we think it is a theoretically superior alternative to the widely used information gain (aka mutual information) function. The results have confirmed the overall superiority of gain ratio over information gain and chi-square when used as a STW function.

Although not proving consistently superior to $tfidf$, STW has given several interesting results. In particular, a STW technique based on gain ratio has given very good results across the board, showing an improvement of 11% over $tfidf$ in macroaveraging for SVMs, currently the best performing TC method in the literature.

6. ACKNOWLEDGEMENTS

We thank Luigi Galavotti for making available his REAL-CAT software [3], with which most of these experiments were performed. A similar thank goes to Thorsten Joachims for making available the SVMLIGHT package [4]. Thanks also to Henri Avancini, Pio Nardiello, Guido Ricci, and Alessandro Sperduti for many fruitful discussions.

7. REFERENCES

[1] T. M. Cover and J. A. Thomas. *Elements of information theory.* John Wiley & Sons, New York, US, 1991.

[2] F. Debole and F. Sebastiani. Supervised term weighting for automated text categorization. Technical Report 2002-TR-08, Istituto di Scienza e Tecnologie dell'Informazione, Consiglio Nazionale delle Ricerche, Pisa, IT, 2002. Submitted for publication.

[3] L. Galavotti, F. Sebastiani, and M. Simi. Experiments on the use of feature selection and negative evidence in automated text categorization. In J. L. Borbinha and T. Baker, editors, *Proceedings of ECDL-00, 4th European Conference on Research and Advanced Technology for Digital Libraries*, pages 59–68, Lisbon, PT, 2000. Springer Verlag, Heidelberg, DE. Published in the "Lecture Notes in Computer Science" series, number 1923.

[4] T. Joachims. Making large-scale SVM learning practical. In B. Schölkopf, C. J. Burges, and A. J. Smola, editors, *Advances in Kernel Methods – Support Vector Learning*, chapter 11, pages 169–184. The MIT Press, Cambridge, US, 1999.

[5] D. D. Lewis. *Representation and learning in information retrieval.* PhD thesis, Department of Computer Science, University of Massachusetts, Amherst, US, 1992.

[6] D. D. Lewis. Evaluating and optmizing autonomous text classification systems. In E. A. Fox, P. Ingwersen, and R. Fidel, editors, *Proceedings of SIGIR-95, 18th ACM International Conference on Research and Development in Information Retrieval*, pages 246–254, Seattle, US, 1995. ACM Press, New York, US.

[7] C. Manning and H. Schütze. *Foundations of Statistical Natural Language Processing.* The MIT Press, Cambridge, US, 1999.

[8] J. R. Quinlan. Induction of decision trees. *Machine Learning*, 1(1):81–106, 1986.

[9] G. Salton and C. Buckley. Term-weighting approaches in automatic text retrieval. *Information Processing and Management*, 24(5):513–523, 1988. Also reprinted in [11], pp. 323–328.

[10] F. Sebastiani. Machine learning in automated text categorization. *ACM Computing Surveys*, 34(1):1–47, 2002.

[11] K. Sparck Jones and P. Willett, editors. *Readings in information retrieval.* Morgan Kaufmann, San Mateo, US, 1997.

[12] Y. Yang and J. O. Pedersen. A comparative study on feature selection in text categorization. In D. H. Fisher, editor, *Proceedings of ICML-97, 14th International Conference on Machine Learning*, pages 412–420, Nashville, US, 1997. Morgan Kaufmann Publishers, San Francisco, US.

[13] J. Zobel and A. Moffat. Exploring the similarity space. *SIGIR Forum*, 32(1):18–34, 1998.