Two novel term weighting for text categorization

L. A. Matsunaga[1] & N. F. F. Ebecken[2]
[1]*Federal District Legislative Assembly, Brazil*
[2]*COPPE/Federal University of Rio de Janeiro, Brazil*

Abstract

In text categorization (TC) based on the vector space model, documents are represented as a vector, where each component is associated with a particular term from the text collection vocabulary. Traditionally, each component value is assigned using the information retrieval (IR) TFIDF measure. While this weighting method seems very appropriate for IR, weighting methods that take into account the importance of the term to the discrimination of the categories may provide better results in TC. To apply this idea, we use in this work variants of TFIDF weighting, where the *idf* part is replaced by functions used to conduct term selection.

In an approach on real-world data to automatically distribute the legislative bills to the committees at the Federal District Legislative Assembly in Brasília, Brazil, the replacement of the *idf* part in TFIDF by a new term selection measure – absl-logit – and by bi-normal separation [1] produced the best general classification results with support vector machines (SVM), when compared with TFIDF and with the use of common term selection measures – chi-square, information gain, gain ratio and odds ratio – to replace the *idf* part in TFIDF.

Keywords: term weighting, text categorization, text classification.

1 Introduction

Text categorization (TC) is the task of automatically assigning unlabelled documents into predefined categories. In TC based on the vector space model, a document is represented as a vector $\mathbf{d}_i^t = [w_{i1}, ..., w_{ip}]$, where p is the size of the text collection vocabulary (number of terms of the dictionary of terms used).

WIT Transactions on Information and Communication Technologies, Vol 40, © 2008 WIT Press
www.witpress.com, ISSN 1743-3517 (on-line)
doi:10.2495/DATA080111

The text collection vocabulary, the dictionary of terms of the problem, is built from the terms used in at least one document of the training set. There are two approaches for the dictionary construction [2]:

- Global dictionary: dictionary composed of the terms used in at least one document of the training set in all categories.
- Local Dictionary: one dictionary is generated per category. Each dictionary is composed of the terms used only in the documents of the training set that form the respective category.

The values w_{ij}, $i=1,\ldots, n$, $j=1,\ldots, p$, between zero and one, represent how much the term t_j contributes to the semantics of document $\mathbf{d}_i$. The most common weighting method used for the weights w_{ij} is TFIDF. There are many variants of TFIDF. The following common variant was used in our experiments, [3, 4]:

$$TFIDF(t_j,\mathbf{d}_i)=\begin{cases}(1+\log f_{ij})\log\dfrac{n}{n_j} & \text{if} \quad f_{ij}>0\\ 0 & \text{otherwise,}\end{cases} \tag{1}$$

Where:

- n: total nº of documents in the training set
- n_j: nº of documents in the training set with term t_j
- f_{ij}: frequency of occurrence of term t_j in document $\mathbf{d}_i$.

Normalization to unit length is generally applied to the resulting vectors.

To improve classification results, Debole and Sebastiani [3] introduced the concept of supervised term weighting (STW), where the information on the membership of training documents into categories is taken into account in the term weighting calculations. Their idea is to replace the *idf* part of TFIDF by a function used to conduct term selection. In their study, the best term weighting performers were obtained using gain ratio and chi-square with global dictionary, SVMs and F_1 macro-averaging.

Motivated by this study, we introduce in this paper two new term weighting methods. In one of these methods, the *idf* part is replaced by a new feature selection method – absl-logit – and in the other one, the *idf* part is replaced by bi-normal separation. The new feature selection method - absl-logit - corresponds to a transformation of the metric odds ratio that corrects the asymmetry presented by this last measure that favors the selection of the more prevalent terms in the positive training examples than in the negative training examples. Bi-normal separation is a term selection method that produced very good results in the study conducted by Forman [1].

In the STW approach, when the local dictionary is adopted, the *idf* value is replaced by the score ($f(t_j,c_k)$) obtained for the term t_j with the term selection method calculated for the category c_k represented by the dictionary. When the global dictionary is adopted, in order to assess the score of a term t_j in a "global" – category independent – sense, it is necessary to use a global measure that summarizes the scores calculated for the term t_j in the individual categories.

The global measure used in this work is the one that produced the best results in previous work [3, 4, 5]:

$$f_{\max}(t_j) = \max_{k=1,\cdots,m} f(t_j, c_k) \tag{2}$$

2 Term selection methods

Term selection is conducted to select the most relevant d terms from the dictionary of terms adopted for the classification task. Before we list the term selection methods that we considered, we introduce some notation.

Table 1 shows the distribution of observed frequencies for term t_j and category c_k.

Table 1: Two-way contingency table of term t_j and category c_k.

category / term	c_k	$c_{\bar{k}}$	total
t_j	n_{kj}	$n_{\bar{k}j}$	n_j
$t_{\bar{j}}$	$n_{k\bar{j}}$	$n_{\bar{k}\bar{j}}$	$n_{\bar{j}}$
total	n_k	$n_{\bar{k}}$	n

n_{kj} : nº of documents in category c_k with term t_j

$n_{k\bar{j}}$: nº of documents in category c_k without term t_j

$n_{\bar{k}j}$: nº of documents not in category c_k with term t_j

$n_{\bar{k}\bar{j}}$: nº of documents not in category c_k without term t_j

n_j : nº of documents in the training set with term t_j

n_k : nº of documents in category c_k

n : total nº of documents in the training set

2.1 Common used term selection methods

Document frequency: measures in how many documents of the training set the term appears;

Chi-square: measures the divergence between the distribution of observed frequencies and the distribution of expected frequencies if one assumes the term t_j occurrence is independent of the category c_k. If term t_j and category c_k are independent, $\chi^2(t_j, c_k)$ is equal to zero**.** The larger the chi-square value, the stronger the association between term t_j and category c_k. The measure is defined to be:

$$\chi^2(t_j, c_k) = \frac{n(n_{kj}n_{\bar{k}\bar{j}} - n_{\bar{k}j}n_{k\bar{j}})^2}{n_k n_j n_{\bar{k}} n_{\bar{j}}}. \tag{3}$$

Information gain: measures the number of bits of information obtained for category c_k prediction by knowing the presence or absence of a term t_j in a document. The larger the information gain value, the more informative is the term t_j for the prediction of the category c_k. The measure is defined to be:

$$IG(t_j, c_k) = -\sum_{c \in \{c_k, \bar{c}_k\}} P(c) \log_2 P(c) + P(t_j) \sum_{c \in \{c_k, \bar{c}_k\}} P(c/t_j) \log_2 P(c/t_j)$$

$$+ P(\bar{t}_j) \sum_{c \in \{c_k, \bar{c}_k\}} P(c/\bar{t}_j) \log_2 P(c/\bar{t}_j). \tag{4}$$

Gain ratio: Debole and Sebastiani [3] define the measure as the ratio between the information gain for category c_k and term t_j, and the entropy of category c_k. That is:

$$GR(c_k) = \frac{IG(c_k)}{-\sum_{c \in \{c_k, \bar{c}_k\}} P(c) \log_2 P(c)}. \tag{5}$$

The larger the gain ratio value, the more informative is the term t_j for the prediction of the category c_k.

Bi-normal separation: Forman [1] defines bi-normal separation as:

$$BNS(t_j, c_k) = \left| \Phi^{-1}\left(\frac{n_{kj}}{n_k}\right) - \Phi^{-1}\left(\frac{n_{\bar{k}j}}{n_{\bar{k}}}\right) \right|, \tag{6}$$

where Φ is the standard normal distribution and Φ^{-1} its corresponding inverse. The larger the BNS value, the larger the indication of difference between the prevalences of term t_j in categories c_k and $c_{\bar{k}}$. To avoid numerical problems, $\Phi^{-1}(0)$ is set to be equal to 0.0005.

Odds ratio: measures the odds of term t_j occurring in documents in category c_k divided by the odds of term t_j not occurring in documents in category c_k. The larger the odds ratio value, the larger the odds of term t_j occurring in documents in category c_k. Mladenic and Grobelnik [6] find this to be the best term selection metric among eleven metrics for a Naive Bayes classifier. For category c_k and term t_j, the odds ratio is given by:

$$OR(t_j, c_k) = \frac{P(t_j/c_k) \big/ (1 - P(t_j/c_k))}{P(t_j/\bar{c}_k) \big/ (1 - P(t_j/\bar{c}_k))}. \tag{7}$$

When:

- $OR(t_j, c_k) > 1$, the odds of term t_j occurring in documents in category c_k is greater than the odds of term t_j not occurring in documents in category c_k;

- $OR(t_j, c_k) = 1$, the odds of term t_j occurring in documents in category c_k is the same as the odds of term t_j not occurring in documents in category c_k;
- $OR(t_j, c_k) < 1$, the odds of term t_j occurring in documents in category c_k is smaller than the odds of term t_j not occurring in documents in category c_k;

For category c_k and term t_j, the odds ratio is estimated by:

$$\hat{OR}(t_j, c_k) = \frac{(n_{kj} + 0.5)(n_{\bar{k}\bar{j}} + 0.5)}{(n_{\bar{k}j} + 0.5)(n_{k\bar{j}} + 0.5)}, \tag{8}$$

where the constant 0.5 is added to each observed frequency of the contingence table 1 to avoid numerical problems [8].

2.2 New feature selection method

Abs-logit: the measure is defined as:

$$ABSL(t_j, c_k) = \left|\ln(OR(t_j, c_k))\right| \tag{9}$$

The larger the value of abs-logit, the more different is the odds of term t_j occurring and not occurring in documents in categories c_k.

From the description of odds ratio, we can verify that when the odds of term t_j occurring in documents in category c_k is greater than the odds of term t_j not occurring in documents of category c_k, $OR(t_j, c_k)$ can vary from 1.001 to infinity, while for the case when the odds of term t_j not occurring in documents in category c_k is greater than the odds of term t_j occurring in documents of category c_k, $OR(t_j, c_k)$ can vary only form 0 to 0.999 (considering three decimal places).

This asymmetry is a drawback to using the odds ratio as a measure of the strength of relationship between terms. However, the problem may be solved by applying the logarithmic transformation (log base e) to the odds ratio, getting this way a symmetric measure of association that is known as logit.

As the interest in text categorization is in terms that are distributed more differently in categories c_k and $c_{\bar{k}}$, no matter the term is more prevalent in category c_k or $c_{\bar{k}}$, the more adequate measure for this purpose is the use of the absolute value of the logit, as defined in (9).

3 Classification algorithm

K-nearest neighbors (KNN) and support vector machines (SVM) are two machine learning approaches to text categorization that have shown better performance than other algorithms in previous studies [1, 3, 4, 8, 9]. As KNN algorithm presents a high computational expense in classification time and

cannot be seen as an alternative to SVM in practical applications [8–10], our experiments were conducted only with SVM approach.

4 Dataset

The dataset consists of a text collection of 1.014 legislative bills introduced at the Federal District Legislative Assembly in Brasília, Brazil, in 2003–2004 [11]. These bills are distributed to the committees for analysis according to the matters within theirs respective jurisdictions.

In the application of automated text categorization to this problem, the documents correspond to the bills to be distributed and the categories correspond to the committees adequate to analyze them.

At the Federal District Legislative Assembly in Brasília, Brazil, there are nine committees, but as the Constitution and Justice Committee analyzes all the bills, we considered in this study only the other remaining eight committees, which are: 1) Economy, Budget and Finance (590 examples); 2) Social Affairs (454 examples); 3) Consumer Protection (102 examples); 4) Human Rights, Citizenship, Ethic and Representative's Decorum (56 examples); 5) Ground Affairs (113 examples); 6) Education and Health (205 examples); 7) Security (148 examples); 8) Economic Development, Science, Technology, Environment and Tourism (189 examples).

From this total (1.014 bills), 355 were analyzed by only one committee, 486 by two committees, 162 by three committees and 11 by four committees. The mean number of committees that analyzed the bills is 1,83 with standard deviation of 0,72.

5 Experimental settings and results

5.1 Experimental settings

Several studies were conducted to choose the best vector representation for the documents. The following aspects of the problem were considered:

1. depending on the matter, the bill may be analyzed by more than one committee, referring the studied problem to a multilabel classification problem. In our experiments, we divided each categorization task into k independent binary classification problems, as usual;
2. a bill (document to be classified) is composed essentially by a summary containing the goal of the law, by the matter to be legislated by the law and by a number of arguments justifying the need for the law. When well written, this summary gives a good indication of the content of the law and should contain the most important words of the text. So, in order to improve classification results, we studied counting twice the frequency of occurrence of the terms used in the summary of the bill. Similar idea was used by Apté et al. [2] for the one-line headline in the Reuters newswire stories;
3. in general, in text categorization problems the categories are viewed just as symbolic labels and no additional knowledge of their meanings are

available. Nevertheless, in the case studied, there is a legislation that indicates the matters under each committee's jurisdiction. To try to improve classification results, we selected, in the bills of the training set, the most important terms related to the committees' jurisdiction and studied increasing the weights for these terms. The idea of increasing the weights for some important words was also used by Schweighofer et al. [12];

4. we have considered dimension reduction only by document frequency. As observed by Debole and Sebastiani [3] and confirmed in a pilot study of this work, dimension reduction using the same term selection function as the one used in the term weighting calculations provides worse classification performance than with the complete dimension;
5. we have studied the use of local and global dictionaries;
6. in our experiments, we used the support vector machines classification algorithm as implemented in the SVM_{light} package [13]. We used the default parameters of the package (among them, linear kernel);
7. as performance measure, we considered F_1-measure (the harmonic mean of the precision and recall). There are two methods for averaging the F_1-measure over a collection of 2-class classification problems. One is the *macro-averaged F_1-measure*, which is the traditional arithmetic mean of the F_1-measure computed for each problem. Another is the *micro-averaged F_1-measure*, which is an average weighted by the category distribution. The former gives equal weight to each problem while the latter gives equal weight to each document classification. Since highly skewed – small categories – problems tend to be more difficult, the macro-averaged F_1-measure tends to be lower. We focused on macro-averaging because we were interested in giving equal weight to each classification problem;
8. to measure performance for a given vector representation studied, we used 5-fold stratified cross-validation;
9. The vector (vi) used to produce all the experiments related in this paper is composed of 14.795 distinct terms and was obtained in a pilot study after: 1) removing the unwanted symbols; 2) removing the terms that appeared in only one bill of the training set; 2) removing the terms that did not contribute to the semantics of a document in the training set (stop words); 3) standardizing the terms written in more than one way.

5.2 Results

Figures 1 and 2 show the comparison of the performance of the initial vector (vi) with the best result or combinations of the best results obtained in each conducted study. In the study of dimension reduction by document frequency (removing those terms whose document frequency were less than two to less than 25), the best result was attained by removing the terms that appeared in less than nine documents of the training set (f9). In the study of increasing the weight of important terms related to the matters within the committees' jurisdiction (considering increases of 30%, 50%, 70% and 100%), the best result was attained by increasing in 30% the weight for these terms (c3). Finally, counting

twice the frequency of occurrence of the terms used in the summary of the bill (s2) also produced good results when compared with the initial vector (vi).

In figures 1 and 2:

- the vector representation 1 is the initial vector (vi);

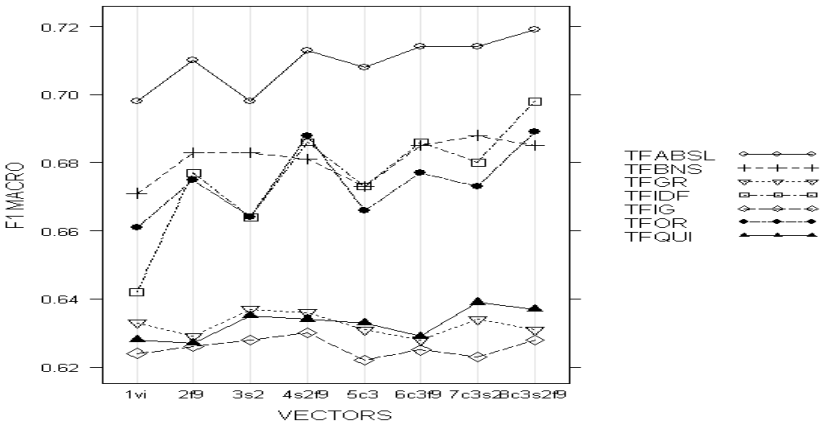

Figure 1: Comparison of the best vector representations with the initial vector (vi), using the global dictionary.

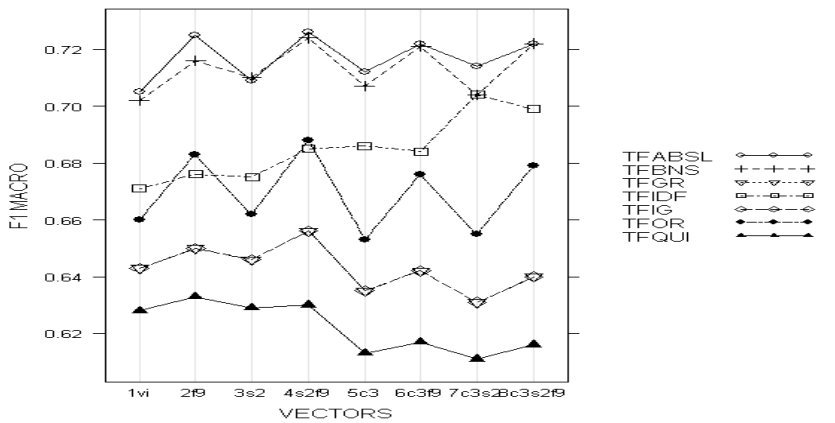

Figure 2: Comparison of the best vector representations with the initial vector (vi), using the local dictionary.

- the vector representation 2 is formed by the terms that occurred in nine or more documents of the adopted dictionary;
- the vector representation 3 is the vector formed by counting twice the frequency of occurrence of the terms used in the summary of the bills;

- the vector representation 4 is formed by the terms that occurred in nine or more documents of the adopted dictionary and by counting twice the frequency of occurrence of the terms used in the summary of the bills;
- the vector representation 5 is the vector formed by increasing the weight of important terms related to the matters within the committees' jurisdiction in 30%;
- the vector representation 6 is the vector formed by the terms that occurred in nine or more documents of the adopted dictionary, increasing the weight of important terms related to the matters within the committees' jurisdiction in 30%;
- the vector representation 7 is the vector formed by counting twice the frequency of occurrence of the terms used in the summary of the bills and by increasing the weight of important terms related to the matters within the committees' jurisdiction in 30%;
- the vector representation 8 is the vector formed by the terms that occurred in nine or more documents of the adopted dictionary, counts twice the frequency of occurrence of the terms used in the summary of the bills and increases the weight of important terms related to the matters within the committees' jurisdiction in 30%.

From figs. 1 and 2, we can verify that the greatest F_1 macro value is achieved by the vector representation 4, fig 2, that uses local dictionary, considers only the terms that appear in at least nine bills (f9), emphasizes the terms used in the summary of the bills by counting their frequency of occurrence twice (s2), and uses term weighting calculated by TFABSL, i.e. replaces the *idf* component in the TFIDF formula by the new term selection method (absl-logit).

We can also observe from fig 1 that with global dictionary, in all conducted studies, the best performances were attained by the weighting method TFABSL. From fig 2, we still can see that with local dictionary, the best performance results are in general attained by TFABSL, with the weighting method TFBNS – i.e. using bi-normal separation to replace the *idf* component in the TFIDF formula – giving very similar results.

6 Final consideration

To confirm the excellent performance of the weighting methods TFABSL and TFBNS proposed in this work, we intend to investigate these methods in standard benchmarks like the Reuters collection and the 20 Newsgroups corpus.

References

[1] Forman, G., An extensive empirical study of feature selection metrics for text classification, *Journal of Machine Learning Research*, 3, pp. 1289–1305, 2003.

[2] Apté, C., Damerau, F. & Wess, S. H., Automated learning of decision rules for text categorization, *ACM Transactions on Information Systems*, **12(3)**, pp. 233–251, 1994.
[3] Debole, F. & Sebastiani, F., Supervised Term Weighting for Automated Text Categorization. *Proc. of the 18th ACM Symposium On Applied Computing*, Melbourne, pp. 784–788, 2003.
[4] Lewis, D. D., Yang, Y., Rose, T. et al., RCV1: A new benchmark collection for text categorization research. *Journal of Machine Learning Research*, **5**, pp. 361–397, 2004.
[5] Yang, Y. & Pedersen, J.O., A comparative study on feature selection in text categorization. *Proc. of the 14th Int. Conf. On Machine Learning*, Nashville, pp. 412–420, 1997.
[6] Mladenic, D. & Grobelnik, M., Feature Selection for unbalanced class distribution and Naïve Bayes. *Proc. of the 16th Int. Conf. On Machine Learning*, pp. 258–267, 1999
[7] Agresti, A., On the logit confidence intervals for the odds ratio with small samples, *Biometrics*, **55(2)**, pp. 597–602, 1999.
[8] Leopold, E. & Kindermann, J., Text categorization with support vector machines. How to represent texts in input spaces?. *Machine Learning*, **46(1–3)**, pp. 423–444, 2002.
[9] Lan, M., Tan, C. L., Low, H. B., Proposing a New Term Weighting Scheme for Text Categorization. *Proc. of the 21st National Conference On Artificial Intelligence*, Boston, pp. 763–768, 2006.
[10] Debole, F. & Sebastiani, F., An analysis of the relative hardness of Reuters-21578 Subsets, *Journal of the American Society for Information Science and Technology*, **56(6)**, pp. 584–596, 2005.
[11] Matsunaga, L., An automated text categorization methodology to distribute the bills to the committees at the Federal District Legislative Assembly. Dept of Civil Engineering, COPPE/Federal University of Rio de Janeiro, 2007.
[12] Schweighofer, E., Haneder, G., Rauber, A. et al., Improvement of Vector Representations of Legal documents with Legal Ontologies. *Proc. of the 5th Int. Conf. On Business Information Systems*, Poznan, 2002.
[13] Joachims, T., Making large-Scale SVM *Learning Practical. Advances in Kernel Methods - Support Vector Learning*, B. Schölkopf and C. Burges and A. Smola (ed.), MIT-Press, 1999.