Experiments in Text Categorization Using Term Selection by Distance to Transition Point

Edgar Moyotl-Hernández, Héctor Jiménez-Salazar

Facultad de Ciencias de la Computación,
B. Universidad Autónoma de Puebla,
14 Sur y Av. San Claudio. Edif. 135. Ciudad Universitaria,
Puebla, Pue. 72570. México,
Tel. (01222) 229 55 00 ext. 7212 Fax (01222) 229 56 72,
emoyotl@mail.cs.buap.mx, hjimenez@fcfm.buap.mx

Abstract. This paper presents a novel term selection method called *distance to transition point* (DTP) that is equally effective for unsupervised and supervised term selection. DTP computes the distance between the frequency of a term and the *transition point* (TP) and then, by using this distance as a criterion, it selects the terms more close to TP. Experimental results on Spanish texts show that feature selection by DTP achieves superior performance to *document frequency*, and comparable performance to *information gain* and *chi-statistic*. Moreover, when DTP is used to select terms in an unsupervised policy, it improves the performance of traditional classification algorithms such as *k*-NN and Rocchio.

Keywords: distance to transition point, term selection, text categorization.

1 Introduction

The rapid growth in the volume of text documents available electronically has led to an increased interest in developing tools that allow organize textual information. *Text categorization* (TC), which is the classification of text documents into a set of predefined categories, is an important task for handling and organizing textual information. Since building text classifiers manually is difficult and time consuming, the dominant approach to TC is based on *machine learning* techniques [10]. Within this approach, a classification learning algorithm automatically builds a text classifier from a set of preclassified documents, a *training set*.

In TC a document d_j is usually represented as a vector of term weights $d_j=(w_{1j},...,w_{Vj})$, where V is the number of terms (the vocabulary size) that occur in the training set, and w_{ij} measures the importance of term t_i for the characterization of document d_j. However, many classification algorithms are computationally hard, and their computational cost is a function of V [2]. Hence, *feature selection* (FS) techniques are used to select a subset from the original term set in order to improve categorization effectiveness and reduce computational complexity. In [12] five FS meth-

ods were tested: *document frequency*, *information gain*, *chi-statistic*, *mutual information* and *term strength*. The first three were found the most effective. For that reason they will be tested in this paper.

A widely used approach to FS is the *filtering*, which consist in selecting the terms that score highest according to a criterion that measures the importance of the term for the TC task [4]. There are two main policies to perform term selection: an *unsupervised* policy, where term scores are determined without using any category information, and a *supervised* policy, where information on the membership of training documents is used to determine term scores [5].

In this paper we present a new term selection method called *distance to transition point* (DTP), which can be used for both unsupervised and supervised term selection. DTP computes the distance between the frequency of a term and the *transition point* (TP), i.e., the frequency that splits the terms of a text (or a set of texts) into low frequency terms and high frequency terms. In the case of unsupervised policy, DTP calculates TP using all training documents, whereas in the case of supervised policy, DTP calculates TP using the training documents belonging to a specific category. We report experimental results obtained on Spanish texts with two classification algorithms: *k*-NN and Rocchio, three term selection techniques: *document frequency* (DF), *information gain* (IG) and *chi-statistic* (CHI), and both unsupervised and supervised term selection by DTP.

The paper is organized as follows. Section 2 briefly introduces the term selection methods (DF, IG and CHI). Section 3 presents the details of the DTP term selection method for both unsupervised and supervised policies. Section 4 describes the classifiers and data used in the experiments. Section 5 presents our experiments and results. Section 6 concludes.

2 Term Selection Methods

In this section we give a brief introduction on three effective FS techniques, one unsupervised method (document frequency) and two supervised methods (information gain and chi-statistic). These methods assign a score to each term and then select the terms that score highest. In the following, let D be the training set, N the number of documents in D, V the number of terms in D, and $C=\{c_1,...,c_M\}$ the set of categories.

Document Frequency (DF). The document frequency of a term t_i is the number of documents in which this term occurs [9]. DF is a traditional term selection method that does not need the category information. It is the simplest technique and easily scales to a large data set with a computation complexity approximately linear in the number N [12].

Information Gain (IG). Information gain of a term t_i measures the number of bits of information obtained by knowing the presence or absence of t_i in a document. If t_i occurs equally frequently in all categories, then its IG is 0. The information gain of term t_i is defined as

$$IG(t_i) = -\sum_{k=1}^{M} P(c_k)\log P(c_k) + P(t_i)\sum_{k=1}^{M} P(c_k \mid t_i)\log P(c_k \mid t_i) + P(\bar{t}_i)\sum_{k=1}^{M} P(c_k \mid \bar{t}_i)\log P(c_k \mid \bar{t}_i) \quad (1)$$

where $P(c_k)$ is the number of documents belonging to category c_k divided by N, $P(t_i)$ is the number of documents with term t_i divided by N, $P(c_k|t_i)$ is the number of documents belonging to c_k with t_i divided by the total number of documents with t_i. The computation includes the estimation of the conditional probabilities of a category given a term, and the entropy computations in the definition. The probability estimation has a time complexity of $O(N)$ and the entropy computations has a time complexity of $O(VM)$ [12].

Chi-Statistic (CHI). The chi-statistic method measures the lack of independence between the term and the category. If term t_i and category c_k are independent, then CHI is 0. In TC, given a two-way contingency table for each term t_i and category c_k (as represented in Table 1), CHI is calculated as follows

$$CHI(t_i, c_k) = \frac{N(ad-cb)^2}{(a+c)(b+d)(a+b)(c+d)} \quad (2)$$

where a, b, c and d are the number of documents for each combination of c_k, $\bar{c}_k$ and t_i, $\bar{t}_i$. In order to get a global score $CHI(t_i)$ from $CHI(t_i, c_k)$ scores relative to the individual categories, the maximum score $CHI_{\max}(t_i) = \max_{k=1}^{M}\{CHI(t_i, c_k)\}$ is used. The computation of CHI scores has a quadratic complexity, similar to IG [12].

Table 1. Two-way contingency table

Category/Term	t_i	$\bar{t}_i$
c_k	a	b
$\bar{c}_k$	c	d

Yang and Pedersen [12] have shown that IG and CHI are the most effective FS methods for k-NN and LLSF classification algorithms. Term selection based on DF had similar performance to IG and CHI methods. The latter result seems to states that the most important terms for categorization are those that occur more frequently in the training set.

3 Distance to Transition Point

Our term selection method DTP is based on TP. TP is derived from the *Law of Zipf* [1],[11],[14], and is the frequency that splits the terms of a text (or a set of texts) into low frequency terms and high frequency terms. In [11] it was observed that TP indicates the frequency around which there are *key words* of a text. In our previous experiments [7] we found that performance of categorization can be slightly increased if terms that occur more often than TP are disregarded. In this paper TP is used to measure the importance of the term for the categorization task. Such measure is an inverse function of the distance between the frequency of a term and the TP; when the frequency of a term is identical to TP, the distance will be zero, producing a maximum closeness score. Throughout the rest of this section we describe the computation of TP and the details of DTP for both unsupervised and supervised policies.

The computation of TP is performed as follows. Let T be a text (or a set of texts), and let I_1 be the number of terms with frequency 1. Then according to [11] the *transition point* of T is defined as

$$TP = (\sqrt{1+8I_1} - 1)/2 \tag{3}$$

As we can see, TP calculation only requires scanning the vocabulary of T in order to find I_1 (for more details on TP see [11] and [8]).

DTP unsupervised. DTP computes the distance to TP in the unsupervised policy as follows

$$DTP(t_i) = |TP - frq(t_i)| \tag{4}$$

where $frq(t_i)$ is the frequency of t_i in D (D is the training set) and TP is computed on D. The computation has a time complexity of $O(V)$.

DTP supervised. In the case of supervised term selection, DTP uses the category information

$$DTP(t_i, c_k) = |TP_k - frq_k(t_i)| \tag{5}$$

where $frq_k(t_i)$ is the frequency of t_i in D_k (D_k is the set of training documents belonging to a specific category c_k) and TP_k is computed on D_k. As the globalization technique we have chosen DTP_{max} because, in preliminary experiments [8], it consistently outperformed other globalization techniques. The computation includes the calculation of the TP for each category and has a time complexity of $O(VM)$.

DTP (whose use as a FS function was first proposed in [8]) selects the terms more close to TP. In FS we measure how close the frequency of a term and TP are to eachother. Thus the terms with the highest value for DTP are the more distant to TP; since we are interested in the terms less distant, we select the terms for which DTP is lowest. Our experiments presented in Section 5 show that the performance of traditional classification algorithms (such as k-NN and Rocchio) is outperformed by term selection with DTP.

4 Classifiers and Data

In order to assess the effectiveness of FS methods we used two classifiers frequently used as a baseline in TC, *k*-NN [13] and Rocchio [3], both treat documents as term vectors.

k-NN is based on the categories assigned to the *k* nearest training documents to the new document. The categories of these neighbors are weighted using the similarity of each neighbor to the new document, where the similarity is measured by the cosine between the two document vectors. If one category belongs to multiple neighbors, then the sum of the similarity scores of these neighbors is the weight of the category [2],[10],[13]. Rocchio is based on the relevance feedback algorithm originally proposed for information retrieval. The basic idea is to construct a prototype vector for each category using a training set of documents. Given a category, the vectors of documents belonging to this category are given a positive weight, and the vectors of remaining documents are given a negative weight. By adding these positively and negatively weighted vectors, the prototype vector of this category is obtained. To classify a new document, the cosine between the new document and prototype vector is computed [6],[10],[13].

The texts used in our experiments are Spanish news downloaded from Mexican newspaper *La Jornada*. We preprocess the texts removing *stopwords*, punctuation and numbers, and stemming the remaining words by means of a Porter's stemmer adapted to Spanish. Term weighting was done by means of the standard *tf·idf* function [9]. We have used a total of 1,449 documents belonging to six different categories (C: Culture, S: Sports, E: Economy, W: World, P: Politics, J: Society & Justice) for training and two testing sets (see Table 2). We only managed one label setting, i.e., each document was assigned in only one category.

Table 2. Training and testing data

	Categories	C	S	E	W	P	J
Training data	No. of documents	104	114	107	127	93	91
	No. of terms	7,205	4,747	3,855	5,922	4,857	4,458
Test data set 1	No. of documents	58	57	69	78	89	56
	No. of terms	5,301	3,333	3,286	4,659	4,708	3,411
Test data set 2	No. of documents	83	65	61	51	90	56
	No. of terms	6,420	3,855	2,831	3,661	4,946	3,822

To evaluate the effectiveness of the classification of documents by classifier, the standard precision, recall and F_1 measures were used. Precision is the number of documents correctly classified, divided by the total number of documents classified. Recall is the number of documents correctly classified, divided by the total number of documents that should be classified. The F_1 measure combines precision (P) and recall (R) as follows: $F_1 = 2RP/(R+P)$. These values can be computed for each individual category first and then be averaged over all categories. Or they can be globally

computed over all the categories. These strategies are respectively called *macroaveraging* and *microaveraged*. Same as [10], we evaluated microaveraged (F_1).

5 Experiments

We performed our FS experiments with both, a k-NN classifier (using $k = 30$), and a Rocchio classifier (where $\beta = 16$ and $\alpha = 4$ as used in [6]). In these experiments we compared three baseline term selection techniques: DF, IG and CHI_{max}, and two variants of our DTP technique: DTP and DTP_{max}. Table 3 lists our F_1 values obtained for k-NN and Rocchio with the evaluated FS techniques at different percent of terms (the vocabulary size in the training set is 14,272).

Table 3. Microaveraged F_1 values for k-NN and Rocchio on test sets

	k-NN					Rocchio				
Percent of terms	DF	IG	CHI_{max}	DTP	DTP_{max}	DF	IG	CHI_{max}	DTP	DTP_{max}
1	.627	.716	.720	.676	.667	.611	.723	.712	.681	.668
3	.697	.769	.758	.759	.756	.701	.756	.749	.742	.739
5	.754	.780	.779	.760	.786	.750	.767	.760	.774	.780
10	.782	.806	.803	.791	.797	.775	.783	.787	.807	.788
15	.802	.811	.801	.807	.799	.782	.801	.793	.811	.791
20	.807	.811	.804	.811	.804	.799	.806	.806	.820	.803
25	.804	.824	.813	.815	.806	.799	.806	.811	.815	.804
50	.809	.813	.803	.814	.806	.807	.807	.815	.829	.811

As seen in table 3, on both k-NN and Rocchio tests DTP is superior to DF, and comparable to IG and CHI_{max} up to percents of terms around 5% and 3% respectively, but becomes superior for percents higher than those. These results, obtained under both DTP variants show that an unsupervised policy performs better than its supervised counterpart.

Results published in [12] showed that common terms are often informative, and viceversa. Our results under DTP do not contradict this for, only the terms that have an extremely low or high frequency are removed, while the terms with *medium* frequency score highest and are preserved. Another interesting result is that DTP unsupervised, while not using category information from the training set, has a performance similar to supervised IG and CHI. In addition to that DTP is much easier to compute than IG and CHI.

6 Conclusions

In this paper we have presented a novel term selection method for TC: *distance to transition point* (DTP), which is based on the proximity to the frequency that splits the terms of a text as low and high frequency terms, i.e., the *transition point* (TP).

Experiments performed on Spanish texts with two classifiers (*k*-NN and Rocchio) showed that feature selection by DTP achieves superior performance to *document frequency*, and comparable performance to *information gain* and *chi-statistic*; three well known and effective techniques. Remarkably, DTP is a simple and easy to compute method.

The degree of enhancement from our method in TC and its relationship to other methods in the literature is the subject of future investigations by the authors.

References

1) Booth, A.: A Law of Occurrences for Words of Low Frequency, Information and Control, (1967) 10(4) 386–93.
2) Galavotti, L., Sebastiani, F., Simi, M.: Experiments on the Use of Feature Selection and Negative Evidence in Automated Text Categorization, Proc. of ECDL-00, 4th European Conference on Research and Advanced Technology for Digital Libraries, (2000) 59–68.
3) Joachims, T.: A Probabilistic Analysis of the Rocchio Algorithm with TFIDF for Text Categorization, Proc. of ICML-97, 14th Int. Conf. on Machine Learning, (1997) 143–151.
4) John, G.H., Kohavi, R., Pfleger, K.: Irrelevant Features and the Subset Selection Problem, Proc. of ICML-94, 11th Int. Conf. on Machine Learning, (1994) 121–129.
5) Karypis, G., Han, E.H.: Concept Indexing: A Fast Dimensionality Reduction Algorithm with Applications to Document Retrieval & Categorization, Technical Report TR-00-0016, University of Minnesota, (2000).
6) Lewis, D.D., Schapire, R.E., Callan, J.P., Papka, R.: Training Algorithms for Linear Text Classifiers, Proc. of SIGIR-96, 19th ACM Int. Conf. on Research and Development in Information Retrieval, (1996) 298–306.
7) Moyotl, E., Jiménez, H.: An Analysis on Frequency of Terms for Text Categorization, Proc. of SEPLN-04, (2004) 141–146.
8) Moyotl, E., Jiménez, H.: Distancia al Punto de Transición: Un Nuevo Método de Selección de Términos para Categorización de Textos, Tesis de Licenciatura, Facultad de Ciencias de la Computación, BUAP, Puebla, México, (2004).
9) Salton, G., Wong, A., Yang, C.: A Vector Space Model for Automatic Indexing, Communications of the ACM, (1975) 18(11) 613–620.
10) Sebastiani, F.: Machine Learning in Automated Text Categorization, ACM Computing Surveys, Vol. 34(1), (2002) 1–47.
11) Urbizagástegui-Alvarado, R.: Las posibilidades de la ley de Zipf en la indización automática, Reporte de la Universidad de California Riverside, (1999).
12) Yang, Y., Pedersen, P.: A Comparative Study on Feature Selection in Text Categorization, Proc. of ICML-97, 14th Int. Conf. on Machine Learning, (1997) 412–420.
13) Yang, Y., Liu, X.: A Re-examination of Text Categorization Methods, Proc. of SIGIR-99, 22nd ACM Int. Conf. on Research and Development in Information Retrieval, (1999) 42–49.
14) Zipf, G.K.: Human Behaviour and the Principle of Least Effort, Addison-Wesley, (1949).