Some Tests in Text Categorization using Term Selection by DTP

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. Distance to Transition Point (DTP) has shown good performance in term selection terms for Text Categorization task. Previous experiment report that DTP behaves as well as CHI and DF term selection techniques. In this paper we present the results of using DTP computed in a global and local fashion; considering the whole of categories of training set. The results confirm that performance of DTP globally computed is better than DTP locally computed. The test carried out took into account two classification methods: k-NN and Rocchio's algorithm; and well known methods to select terms: CHI, IG and DF.
Keywords: distance to transition point, term selection, text categorization.

1 Introduction

We are interested in the automated assignment of texts into predefined categories, Text Categorization (TC), which is solved by supervised learning algorithms [10]. Specifically, our interest is in the selection of terms from texts in order to train supervised algorithms. This is an important problem because of the necessity to optimize time and memory space.

Some classification algorithms used in TC are: Rocchio's algorithm, $k-$Nearest Neighbor (kNN), and Support Vector Machines [6][2][13]. Also there are several methods of term selection [7][12], and techniques to term selection. Most known term selection techniques assign weights to terms, and those which guess more importance are selected. Three very used criteria to term selection are [7][12]: Document Frequency (DF), which assigns to t_i the number of documents that use t_i; χ^2 statistic (CHI), is the value representing the lack of independence between the term and the category; and Information Gain (IG), which measures the number of bits of information by knowing the presence or absence of a term in a document.

Our approach to term selection is based in term frequency in the whole of the collection. It distinguishes a frequency called *Transition Point* (TP) which is used to compute the distance of each term frequency to TP (DTP). Previous

experiments [5] have shown that DTP improves in some aspects other term selection methods as CHI and DF; for example, this due to DTP calculation is an $O(n)$ algorithm where n is the number of terms in collection, and it improves slightly the performance of TC task. Later has motivated to continue the research on DTP.

In this paper we present an experiment varying the calculation of TP in two different ways: local TP, using the documents of each category to calculate TP; and global TP, using the whole of the documents to calculate it. The experiment was accomplished on a small texts collection (6 categories with totally around 1,000 documents), using two classification methods: Rocchio's and k-NN.

The following section explains some backgrounds about DTP. Sections 3, 4 and 5 describe the material and methods used in the test accomplished. At the end, we provide the conclusions reached in this work.

2 Distance to Transition Point

As we have said our method is based on *Transition Point* (TP). TP is the frequency of a term which splits terms into high frequency terms and low frequency terms. The present test is based on a previous experiment that shown that TP behaves well as a cut on the selected term by the classical methods [4]. From the fact that TP indicates the frequency which around it there are key words of the text [11], was calculated a weight per term. Such weight, for term t_i, is calculated as the inverse of distance of t_i frequency to TP (DTP): the more closeness of frequency to TP, the more weight for that term. Firstly, let us see how to calculate TP; some details may be found in [5].

Let T be a text (or a set of texts), and let I_1 be the number of words with frequency 1. TP is defined as [11]:

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

As we can see, TP calculation requires only scanning the full text in order to find I_1, which can be done in $O(N)$, where N is the number of terms.

Now, DTP is easily calculated making the difference between each term frequency and TP. We would hope that DTP calculated for each category ports more information and, therefore, TC were better. However, test accomplished shown that DTP on the basis of all categories has better performance. We will show this fact in the following sections.

3 Term Selection Methods

In this section, we give a brief introduction on two effective term selection (TS) techniques as they are presented in [12], including one unsupervised method (in the sense that it does not use category information) DF and two supervised method (it uses category information) CHI and IG. These methods assign a score to each individual term and then select the terms that score highest. In

the following, let us denote with D the training documents set, N_D the number of documents in the training set, and let $\{c_k\}_{k=1}^{M}$ the categories set. Selection methods used in the test were the following:

Document Frequency (DF). Document frequency is the number of documents in which a term t_i occurs. It is the simplest technique for term selection and easily scales to a large data set with a computation complexity approximately linear in the number N_D. It is a simple but effective term selection method for TC [12].

χ^2 statistic (CHI). The χ^2 statistic measures the lack of independence between the term and the category. In the TC, given a two-way contingency table for each term t_i and category c_k as represented in Table 1, it is defined to be:

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

where, a, b, c and d indicate the number of documents for each cell in the contingency table (see table 1). We computed CHI for each category and each term in D, and then combined the category specific scores of each term into one score: $CHI(t_i) = max_{k=1}^{M}\{CHI(t_i, c_k)\}$. The computation of CHI scores has a quadratic complexity [12].

Information Gain (IG). Information gain of a term measures the number of bits of information obtained by knowing the presence or absence of a term in a document. The information gain of term t_i is defined as

$$\begin{aligned} IG(t_i) = &- \sum_{k=1}^{M} P(c_k) \log P(c_k) \\ &+ P(t_i) \sum_{k=1}^{M} P(c_k|t_i) \log P(c_k|t_i) \\ &+ P(\overline{t_i}) \sum_{k=1}^{M} P(c_k|\overline{t_i}) \log P(c_k|\overline{t_i}) \end{aligned} \tag{3}$$

where, for example, $P(c_k)$ is the number of documents belonging to the category c_k divided by the total number of documents in D_{Tr}, $P(t_i)$ is the number of documents without the term t_i divided by the total number of documents in D_{Tr}, $P(c_k|t_i)$ is the number of documents of category c_k with the term t_i divided by the number of documents with t_i, etc. 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 $O(N)$ an the entropy computations has a time complexity of $O(VM)$ [12].

Distance to Transition Point (DTP). DTP measures importance of a term according to the distance of that term to TP:

$$DTP(t_i) = |TP - \mathrm{fr}(t_i)|, \tag{4}$$

where $\mathrm{fr}(t_i)$ is the frequency of t_i in D, and TP is computed on D.

Category/Term	t_i	$\overline{t_i}$
c_k	a	b
$\overline{c_k}$	c	d

Table 1. Two-way contingency table.

4 Classification Methods

To asses the effectiveness of TS methods we used two classifiers: $k-$NN and Rocchio. Both classifiers treat documents as a feature 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 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. Rocchio is based on the relevance feedback algorithm originally proposed for information retrieval. It has been extensively used for TC. The basic idea is to construct a prototype vector for each category using training 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 summing up 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.

Both classifiers are *context sensitive* in the sense that no independence is assumed between either terms or categories [12]. kNN and Rocchio treat a document as a single point in a vector space, thus enabling a better observation on TS.

5 Test

The texts used in our experiments are Spanish news downloaded from the Mexican newspaper *La Jornada* (year 2000). We preprocess the texts removing *stopwords*, punctuation and numbers, and stemming the remaining words by means of a Porter's stemmer adapted to Spanish. We have used a total of 1,449 documents for training belonging to six different classes: Culture (C), Sports (S), Economy (E), World (W), Politics (P) and Society & Justice (J). Additionaly were used two test data sets (see Table 2). We only managed one label setting (i.e., each document was assigned in only one class).

To evaluate the effectiveness of category assignments to documents by classifier, the standard precision, recall and F_1 measure was used here. Precision is defined to be the number of categories correctly assigned divided by total number of categories assigned. Recall is the number of categories correctly assigned divided by the total number of categories that should be assigned. The F_1 measure combines precision (P) and recall (R) with an equal weight in the following form $F_1 = 2RP/R + P$. There scores can be computed for the binary

	Categories	C	S	E	W	P	J
Training data	No. of documents	104	114	107	127	93	91
	No. of terms	7,131	4,686	3,807	5,860	4,796	4,412
Test data set1	No. of documents	58	57	69	78	89	56
	No. of terms	5,228	3,285	3,235	4,611	4,647	3,774
Test data set2	No. of documents	83	65	61	51	90	56
	No. of terms	6,349	3,799	2,793	3,611	4,879	3,778

Table 2. Training and testing data.

decisions on each individual category first and then be averaged over categories. Or they can be computed globally over all the $N_T \cdot M$ binary decisions where N_T is the number of total test documents, and M is the number of categories in consideration. The former way is called *macroaveraging* and the latter *microaveraging*. We have evaluated microaveraging F_1, since it is almost preferred to macroaveraging [10].

Percent	Number	$k-$NN					Rocchio				
of terms	of terms	DF	CHI	IG	DTP	DTPloc	DF	CHI	IG	DTP	DTPloc
1	142	0.610	0.702	0.710	0.680	0.669	0.616	0.705	0.718	0.690	0.659
3	426	0.696	0.748	0.764	0.763	0.743	0.702	0.738	0.754	0.750	0.764
5	710	0.750	0.776	0.781	0.765	0.744	0.750	0.756	0.761	0.775	0.768
10	1,419	0.791	0.801	0.793	0.791	0.801	0.777	0.781	0.776	0.808	0.783
15	2,129	0.795	0.797	0.802	0.807	0.792	0.777	0.782	0.783	0.812	0.787
20	2,838	0.799	0.798	0.806	0.807	0.802	0.782	0.786	0.790	0.820	0.793
25	3,548	0.801	0.799	0.807	0.809	0.799	0.788	0.795	0.791	0.819	0.792
50	7,095	0.799	0.786	0.803	0.809	0.802	0.795	0.798	0.803	0.822	0.798

Table 3. F_1 average values for $k-$NN and Rocchio on test set1 and set2.

We have performed our TS experiments first with the standar k-NN classifier (with $k = 30$), and subsequently with the Rocchio classifier (with $\beta = 16$ y $\alpha = 4$ as was suggested in [3]). In these experiments we have compared two baseline term selection functions, i.e. DF, CHI, IG and DTP (calculated globally on training collection). Table 3 lists the F_1 values for k-NN and Rocchio with different TS techniques at different percent of terms (the number of different terms in the training set is 14,190).

We calculate the correlation coefficients among the statistics with the formula:

$$\rho_{XY} = \frac{Cov(X,Y)}{\sigma_X \cdot \sigma_Y} \tag{5}$$

$$\tag{6}$$

where X,Y are random variables, $Cov(X,Y)$ is their covariance, and $\sigma_X \cdot \sigma_Y$ are their standard desviations respectively. The results are given in Table 4:

	DF	IG	CHI	DPTglobal	DPTlocal
DF	1	0.573	0.358	0.255	0.443
IG	0.573	1	0.919	0.215	0.361
CHI	0.358	0.919	1	0.167	0.283
DPTglobal	0.255	0.215	0.167	1	0.1
DPTlocal	0.443	0.361	0.283	0.1	1

Table 4. Correlation Coefficients of Statistics.

Yang and Pedersen [12] conducted a comparative study on several TS methods, and found: First, CHI is one of the most effective method to reduce the dimensionality of the term space. Second, DF performance, similarly, was shown scoring in favor of common terms over rare terms. Third, strong correlations between DF and CHI values of a term are general rather than corpus-dependent. The strong correlations means that common terms are often informative, and reciprocally. Such correlation are displayed in table 4. Basically, these results seem to state that the most valuable terms for TC are those with medium frequency in the training set, i.e., those around of the TP. Another interesting conclusion in [12] is that using category information for TS does not seem to be crucial for excellent performance. DTP does not use category information present in the training set, but has a performance similar to CHI.

6 Conclusions

We have showed some features of DTP behavior. First, DTP performance is similar to other term selection techniques, varying two classification methods, Rocchio and $k-$NN. Second, DTP calculation is in the same complexity class that DF, the better technique in execution time. Third, DTP is independent of category, it may be computed globally, and performance of DTP locally computed is a method with low performance in TC task.

Although, the difference between DTP and other term selection techniques is not significative in TC, this result encourages to carry out further experiments to know how much our proposal enhances TC task with respect to reported results in the literature.

References

1. Booth, A.: "A Law of Occurrences for Words of Low Frequency", *Information and control*, 10(4) pp 386-93, 1967.
2. Lam, W. & Ho, C.: "Using Generalized Instance Set for Automatic Texts Categorization", in *Proc. of 21th Ann. Int. ACM SIGIR Conf. on Research and Development in Information Retrieval*, pp 81-89, 1998.
3. Lewis, D.D.: "Training Algorithms for Linear Text Classifiers", *SIGIR96*, 1996.
4. Moyotl, E. & Jiménez, H.: "An Analysis on Frequency of Terms for Text Categorization", to be published in *Proc. of SEPLN*, 2004.

5. Moyotl, E. & Jiménez, H.: "A Novel Method to Select Terms for Text Categorization", sent to *Iberamia 2004*.
6. Rocchio, J.J.: "Relevance Feedback in Information Retrieval". In G. Salton, (Ed.), *The SMART retrieval system: experiments in automatic document processing*, Prentice-Hall, 1971.
7. Rogati, M. & Yang, Y.: "High-Performing Feature Selection for Text Classification", *CIKM'02*, ACM, 2002.
8. Salton, G. & McGill, M.: *Introduction to Modern Information Retrieval*, 1983.
9. Salton, G., Wong, & Yang, C.S.: "A Vector Space Model for Automatic Indexing", *Information Retrieval and Language Processing*, pp 613-620, 1975.
10. Sebastiani, F.: "Machine Learning in Automated Text Categorization", *ACM Computing Surveys*, Vol. 34(1), pp 1-47, 2002.
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", in *Proc. of 14th Int. Conf. on Machine Learning*, pp 412-420, 1997.
13. Yang, Y. & Liu, X.: "A Re-examination of Text Categorization Methods", *ACM SIGIR*, pp 42-49, 1999.
14. Xue, D. & Sun, M: "A Study on Feature Weighting in Chinese Text Categorization", *Lecture Notes in Computer Science*, A. Gelbukh (Ed.), Vol. 2588, Springer, pp 592-601, 2003.
15. Zipf, G.K.: *Human Behaviour and the Principle of Least Effort*, Addison-Wesley, 1949.