

# An Analysis on Frequency of Terms for Text Categorization \*

**Edgar Moyotl-Hernández**

Fac. de Ciencias de la Computación  
B. Universidad Autónoma de Puebla  
C.U. 72570, Puebla, México  
emoyotl@mail.cs.buap.mx,

**Héctor Jiménez-Salazar**

Fac. de Ciencias de la Computación  
B. Universidad Autónoma de Puebla  
C.U. 72570, Puebla, México  
hjimenez@fcfm.buap.mx

**Resumen:** Presentamos resultados sobre una forma de selección de términos con fines de categorización de textos. Usamos el punto de transición, esto es, la frecuencia de un término en un texto o colección de textos que divide en dos a los términos: los de alta y baja frecuencia. Se tomaron porcentajes de términos basados en valores de su frecuencia entre documentos, la ganancia de información y la prueba  $\chi^2$ . El desempeño de la categorización, entrenando el algoritmo de Rocchio con estos términos, fue comparado con un recorte hecho a la cantidad de términos en función de su frecuencia comparada con el punto de transición de la colección de entrenamiento. En un experimento, aplicado a la categorización de textos en español, fueron usados los mencionados criterios de selección y se observó una disminución de términos, manteniendo al menos el mismo desempeño. En nuestro experimento, el mejor desempeño lo obtuvo la selección basada en los valores de frecuencia de términos entre documentos combinada con el punto de transición.

**Palabras clave:** punto de transición, selección de índices, categorización de texto.

**Abstract:** Preliminary results on a way to reduce terms for text categorization are presented. We have used the transition point; a frequency which splits the words of a text into high frequency words and low frequency words. Thresholds outcoming from document frequency of terms, Information Gain and  $\chi^2$  were tested in combination with the transition point. A text categorization experiment based on Rocchio's method showed that selecting terms whose frequency is lesser than the transition point discarded noise terms without diminishing the categorization task performance. In our experiment, the best result was for term selection based on document frequency of terms threshold in combination with the transition point as a cut.

**Keywords:** transition point, term selection, text categorization.

## 1. Introduction

Text Categorization (TC), the automated assignment of texts into predefined categories, is a problem in the supervised learning paradigm (Sebastiani, 2002). Several methods for TC have been proposed as Rocchio's algorithm, nearest neighbor, and support vector machine (Rocchio, 1971)(Lam and Ho, 1998)(Yang and Liu, 1999), among others. The whole of methods require to select the best terms as means to optimize time and memory space (Rogati and Yang, 2002)(Yang and Pedersen, 1997). Besides, term weighting is also an important issue related to feature reduction (Xue and Sun, 2003).

Vector Space Model (VSM) is a framework to weight terms that occur in texts. However, the high volume of features gene-

rated in this model leads to use thresholds applied to the terms' weights in order to diminish the high dimensionality of the terms space. The weights determined in VSM are mainly supported by the number of appearances of a term in the text (term frequency of  $k$ :  $tf_k$ ) and the number of documents that use a term (document frequency:  $df_k$ ). Terms with high  $df_k$  do not help to discern among several texts (due to similar weights of the most texts). Similarly, terms with low  $df_k$  (rare terms) commonly introduce noise when they are considered. G. Salton et. al. (Salton, Wong, and Yang, 1975) empirically showed that terms  $k$  such that  $df_k \notin [m/100, m/10]$  (where  $m$  is the number of documents) are poor terms for the purpose of distinguishing between documents. In contrast, terms which satisfy  $df_k \in [m/100, m/10]$ , called discriminant terms, disperse vectors that represent documents. Thus, using discriminant terms

\* We want to thank the support of VIEP-BUAP through the grant III09G03.

is easier to identify documents. This result permits to reduce the dimension of space up to 25 %. On the other hand, application of thresholds on  $df_k$  achieves vector space dimension reduction near 20 % (Xue and Sun, 2003).

Salton's et. al. empirical fact on  $df_k$  motivates the exploration on medium frequency words. The underlying hypothesis is that terms of medium frequency have usually high semantic content. The accomplished test, in the present work, tries to observe the effect of using discriminant terms in order to make clear which terms are more important in TC task. In our experiment the selection of medium term frequency is faced through the transition point. R. Urbizagástegui (R. Urbizagástegui-Alvarado, 1999) carried out an experiment identifying on a text terms of medium frequency, those around the *transition point* (TP). TP is the frequency of a term which splits terms into high frequency terms and low frequency terms. TP is closely chosen to a frequency for which there is only one term (R. Urbizagástegui-Alvarado, 1999). As we will see, the law for low frequency of words, due to A. Booth (Booth, 1967), is the basis to calculate TP. We will show that the use of TP as a cut criteria may help to discard noise features on term selection.

Then, we present the results of using TP as an additional criteria to the well known techniques of chi-square and information gain thresholding (Yang and Pedersen, 1997)(Rogati and Yang, 2002) applied to terms selection in order to the Spanish texts categorization task.

The following two sections explain some details about discriminant value and transition point background. The fourth section of this work presents the experiment accomplished. At the end, we provide the conclusions reached in this work.

## 2. Discriminant Value

Given a document collection  $\{D_1, \dots, D_m\}$ ,  $D_i$  is represented by a vector  $\vec{D}_i = (d_{i1}, \dots, d_{iN})$  where its components are defined by weights of the terms contained in  $D_i$ :  $d_{ik} = tf_{ik} \cdot idf_k$ , where  $tf_{ik}$  is the term frequency  $k$  into the document  $D_i$ , and  $idf_k$  is defined as:

$$idf_k = \log_2(m) - \log_2(df_k) + 1; \quad (1)$$

$df_k$  denoting the number of documents that contain the term  $k$ . Similarity of two documents is taken as the cosine of the angle between the corresponding vectors.

The discriminant value of a term  $k$ ,  $dv_k$ , is defined as the difference  $Q_k - Q$ , where  $Q$  measures the document density on the full collection. Density is computed using the centroid of the collection,  $\vec{C}$ :

$$Q = \sum_{i=1}^m sim(\vec{C}, \vec{D}_i). \quad (2)$$

The  $Q_k$  value is obtained regarding on vectors all terms except  $k$ . Thus, the higher  $dv_k$  the more density without  $k$  or, the higher  $dv_k$  the higher dispersion with  $k$ . Terms with high  $dv_k$  are preferred because such terms make easier the identification of documents.

Salton et. al. (Salton, Wong, and Yang, 1975) carried out an experiment to identify discriminant terms, i.e. terms with high  $dv$ . Experiment was as follows: discriminant value average on terms with same  $df$  value (*df-class*) was calculated, such values were ranked on  $dv$ , and finally, in order to identify best terms, pairs  $(df, rank_{dv})$  were plotted. As we have said terms  $k$  which satisfy  $df_k \in [m/100, m/10]$  have high  $dv$ .

## 3. Transition Point

Let  $T$  be a text (or a set of texts). Let us consider the list of the words that occur in  $T$  ordered by its frequencies. The top word of this list has rank 1, and successively the words are enumerated. Let us denote with  $p(r)$  the probability that a word has rank  $r$ . If  $T$  has  $N$  words, then a word  $w$  with rank  $r$  and frequency  $f$  should hold:  $Np(r) = f$ . From Zipf's observations (Booth, 1967) it may be stated that if a word holds:

$$2 > Np(r) \geq 1, \quad (3)$$

then such word occurs in the text (or texts) with frequency 1. Zipf's Law suggests that  $p(r) = k/r$ , where  $k$  is a constant linked to the text  $T$ . Then, we have

$$2 > Nk/r \geq 1. \quad (4)$$

From equation 4 we could say that there exist two values, minimum and maximum rank:

$$r_{min} = \frac{1}{2}kN \quad r_{max} = kN, \quad (5)$$

that includes the words with frequency 1. If  $I_1$  denotes the number of words with frequency 1, then  $I_1 = r_{max} - r_{min}$ . This leads to

$$I_1 = \frac{1}{2}kN. \quad (6)$$

The same reasoning used to derive  $I_1$  can be followed for  $I_n$ , the number of words with frequency  $n$ , which gives:

$$I_n = \frac{1}{n(n+1)}kN. \quad (7)$$

From the equations 6 and 7 it is possible to obtain a valid ratio for any text:

$$I_n/I_1 = 2/n(n+1). \quad (8)$$

Now, it is chosen a frequency  $n$  such that  $I_n = 1$ ; because we are interested on medium frequency terms. Thus, from the previous ratio,  $n(n+1) = 2I_1$ , the transition point is defined as (R. Urbizagástegui-Alvarado, 1999)

$$n = (\sqrt{1 + 8I_1} - 1)/2. \quad (9)$$

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.

Using a 25% bandwidth around the transition point, good results on key word identification were reported in (R. Urbizagástegui-Alvarado, 1999).

#### 4. Performed Test

The texts used in our experiments are Spanish news downloaded from the 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 belonging to six different classes (culture (C), sports (S), economy (E), world (W), politics (P) and society & justice (J)) for training and two test data sets (see Table 1). We only used one label setting (i.e., each document was assigned in only one class).

In order to know the behavior of TP on dispersion we carried out a similar  $dv$  computation as Salton et. al.'s one described in sec. 3. In this case we used  $tf_k$  instead of  $df_k$ . Computation was as follows. We calculated  $dv$  value for a sample (1,563 terms) from the whole of training set terms (14,190 terms),

in order to know the influence of TP on text dispersion. The sample covered the 636  $df$ -classes (see section 2). For each term of this sample  $dv$  was determined, in a similar manner as in Salton et. al.'s experiment, the pair  $(tf_k, rank_{dv_k})$ . Such pairs were plotted with a smoothed polygon by a quadratic curve of Bézier (see fig. 1). Smoothing was done to clearly see the curve tendency. The curve shows that terms  $k$  whose  $tf_k \in [10, TP]$  have high  $dv$ . On the basis of this result, TP was used as a cut for some term selection methods.

In order to know the effect of TP as an additional criteria of selection we used three term selection methods (as they are presented in (Yang and Pedersen, 1997)): Document Frequency (DF), Information Gain (IG), and  $\chi^2$  (CHI); maximum value per term was used on CHI. In the following, let us denote the training documents set by  $D$ , the dimension of the feature space by  $V$ , the number of documents in the training set by  $N$ , and let  $\{c_k\}_{k=1}^M$  be the set of categories. Term selection methods are described as follows:

**Document Frequency (DF).** The document frequency of term  $t_i$  is the number of documents in  $D$  in which  $t_i$  occurs. It is a simplest but effective global function for feature selection and easily scales to a large data set with linear computation complexity.

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

$$\begin{aligned} IG_i &= - \sum_{k=1}^M P(c_k) \log P(c_k) \\ &\quad + P(t_i) \sum_{k=1}^M P(c_k|t_i) \log P(c_k|t_i) \\ &\quad + P(\bar{t}_i) \sum_{k=1}^M P(c_k|\bar{t}_i) \log P(c_k|\bar{t}_i) \end{aligned}$$

the probabilities are interpreted on an event space of documents, and are estimated by counting occurrences in the training set.

**$\chi^2$  Statistic (CHI).** It measures the lack of independence between the term and the category. It is defined to be

$$CHI(t_i, c_k) = \frac{N \times (AD - CB)^2}{E} \quad (10)$$

where  $E = (A+C) \times (B+D) \times (A+B) \times (C+D)$ , and  $A$  is the number of documents that contain  $t_i$ ,  $B$  is the number of documents that do not belong to  $c_k$  and have term  $t_i$ ,  $C$  is the number of documents that do not contain  $t_i$ ,  $D$  is the number of documents that do not belong to  $c_k$  and contain  $t_i$ .  $CHI_{ik}$  has value zero if  $t_i$  and  $c_k$  are independent.

We use the TP as cut for term selection. Terms whose value of methods referred above (DF, IG and CHI) is highest are maintained comparing its frequency with TP; it must be lesser than TP.

TC was accomplished using global Rocchio's method on two test sets. Rocchio's method adapted for TC calculates a prototype vector  $\vec{c}_k = (w_{1k}, \dots, w_{Nk})$  for each category, the weight  $w_{ik}$  of  $t_i$  in the category  $c_k$  is a combination of the weight in relevant documents belonging to  $c_k$ , and the weight in non-relevant documents that do not belong to  $c_k$ . This weight is computed as:

$$w_{ik} = \frac{\beta}{|R_k|} \sum_{d_j \in R_k} w_{ij} - \frac{\gamma}{|NR_k|} \sum_{d_j \in NR_k} w_{ij} \quad (11)$$

where  $w_{ij}$  is the weight of  $t_i$  in document  $d_j$ ,  $R_k$  specifies the relevant documents to  $c_k$ , and  $NR_k$  the non-relevant documents to  $c_k$ . In eq. 11  $\beta$  and  $\gamma$  are parameters that control the importance of relevance of documents. We used  $\beta = 16$ ,  $\gamma = 4$ , as suggested in (Lewis, 1996).

To evaluate the effectiveness of category assignments by classifier to documents, the standard precision, recall and  $F_1$  measure are 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  $F_1 = 2RP/R + P$ . These scores can be computed for the binary decisions on each individual category first and then be averaged over categories. Or they can be computed globally over all the  $N \cdot M$  binary decisions where  $N$  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 (Sebastiani, 2002).

We have performed our term selection experiments with the Rocchio classifier. In these experiments we have compared three baseline feature selection methods, i.e. DF, IG and CHI. Table 2 lists the  $F_1$  values for Rocchio with different term selection techniques at different percentages of terms (the number of different terms in the training set is 14,190).

## 5. Conclusions

Table 2 displays the performance, microaverage on  $F_1$ , of the classification task. Tests was done on data set1 and data set2 (see table 1), with Rocchio's algorithm using three term selection methods: DF, IG and CHI (columns 3, 4 and 5 of table 2), and using TP as a cut on terms selected by the same methods, namely: DFcut, IGcut and CHIcut (columns 6, 7 and 8). Terms were ranked according to values given by DF, IG and CHI, and from such lists were taken several percentages (indicated in column 1 of table 2). The number of terms using TP-cut diminished, except for IGcut. Terms selected by IG were not affected by TP-cut, due to the fact that information gain gives higher values to less frequent terms and, therefore, frequency of such terms is lesser than TP. We see slight differences between results of DF and CHI without cut and using TP-cut, respectively; but they have not statistical difference ( $Z = 0,007$  for DF and  $Z = 0,008$  for CHI, using  $p$ -test (Yang and Liu, 1999)). Furthermore,  $F_1$  for DFcut increases more rapidly than CHIcut and uses a lower quantity of terms. We can conclude that selecting terms whose frequency is lesser than TP is a useful criterion in term selection methods for text categorization.

So, TP is a promising criterion for term selection. However, it is still necessary to use several text collections, as Reuters, and different methods of classification, to validate the power of TP in text categorization tasks. An advantage of TP is that it can be easily computed; saving effort to reduce features in combination to other feature selection methods which are, generally, more expensive. Additional test may also help to know contribution of TP and  $dv$  on feature reduction; particularly criteria which take advantage on dispersion into vector space, and properties of informative content of terms.

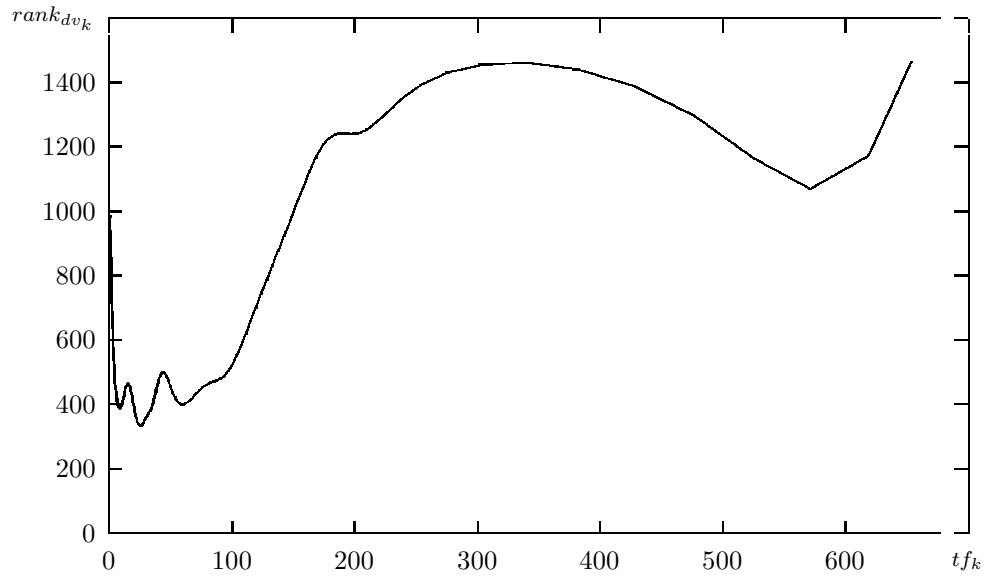


Figura 1: Rank of discriminant values for  $tf_k$ .

	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

Cuadro 1: Training and testing data.

% terms	# terms	DF	IG	CHI	DFcut/#terms	IGcut/#terms	CHIcon/#terms
1	142	0.616	0.491	0.705	0/0	0.491/142	0.622/82
3	426	0.702	0.491	0.738	0.695/274	0.491/426	0.723/366
5	710	0.750	0.487	0.756	0.750/558	0.487/710	0.743/650
10	1,419	0.777	0.551	0.781	0.788/1,267	0.551/1,419	0.775/1,359
15	2,129	0.777	0.633	0.782	0.798/1,977	0.633/2,129	0.779/2,069
20	2,838	0.782	0.624	0.786	0.813/2,686	0.624/2,838	0.787/2,778
25	3,548	0.788	0.622	0.795	0.811/3,396	0.622/3,548	0.795/3,488
50	7,095	0.795	0.752	0.798	0.824/6,943	0.752/7,095	0.804/7,035

Cuadro 2: Microaveraged  $F_1$  for Rocchio based on average of tests for set1 and set2.

## References

- Booth, A. 1967. A Law of Occurrences for Words of Low Frequency. *Information and Control*, 10(4):386–393.
- Lam, W. and C. Ho. 1998. Using Generalized Instance Set for Automatic Texts Categorization. In *Proc. of Conf. on Research and Development in IR*, pages 81–89. ACM.
- Lewis, D. D. 1996. Training Algorithms for Linear Text Classifiers. In *Proc. of Conf. on Research and Development in IR*. ACM.
- R. Urbizagástegui-Alvarado. 1999. Las Posibilidades de la Ley de Zipf. Reporte de la Universidad de California Riverside.
- Rocchio, J. J. 1971. *Relevance Feedback in Information Retrieval*. In G. Salton (Ed.), Prentice-Hall.
- Rogati, M. and Y. Yang. 2002. High-Performing Feature Selection for Text Classification. ACM.
- Salton, G., Wong, and C.S. Yang. 1975. A Vector Space Model for Automatic Indexing. *Information Retrieval and Language Processing*, pages 613–620.
- Sebastiani, F. 2002. Machine Learning in Automated Text Categorization. *ACM Computing Surveys*, 34(1):1–47.
- Xue, D. and M. Sun. 2003. A Study on Feature Weighting in Chinese Text Categorization. *Lecture Notes in Computer Science, A. Gelbukh (Ed.) Springer*, 2588:592–601.
- Yang, Y. and X. Liu. 1999. A Re-examination of Text Categorization Methods. In *Proc. of Conf. on Research and Development in IR*, pages 42–49. ACM.
- Yang, Y. and P. Pedersen. 1997. A Comparative Study on Feature Selection in Text Categorization. In *Proc. of the Fourteenth of the ICML*, pages 412–420.